

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

A.I. vs. Pitfall (by Jay Cordes)

The tagline for this course is: “where bots end, you begin.” This begs the question “where do they end?” The answer can be described succinctly as “the ability to not fool oneself.” To be sure, by default humans are prone to fool themselves, too. But they can learn a scientific mindset that bends over backwards to falsify one’s beliefs and not jump to conclusions based on superficial observations. They can also learn critical thinking and the ability to creatively find ways to drill down to the truth. This course helps data scientists build this muscle, so they can collaborate with bots where they are strong but learn to catch bots and themselves when statistical and cognitive pitfalls arise.

In order to demonstrate where ChatGPT meets its match, we will tackle exercise 1 from the course as a case study. Below is the description we provide to students and after that are the most interesting portions of the chat history in which I tried to guide ChatGPT towards the solution. To give it every advantage, I even created a custom GPT into which I uploaded a copy of my book *The 9 Pitfalls of Data Science*. The pitfall encountered in this exercise is one of the 9 highlighted in the book and had an entire chapter dedicated to it.

The bottom line is that ChatGPT has tons of knowledge, but is unable to think. As I’ll show below, I could lead the ChatGPT horse to the regression towards the mean water, but I couldn’t make it drink. In order to avoid fooling ourselves, we need the ability to suppose we’re wrong and then come up with ways to try to rule it out. There’s no simple way to “be sure” that you’ve got the right idea, so it’s unlikely that following automated instructions or reciting memorized knowledge will ever be enough. You have to think.

The Challenge Given to Students - Exercise 1: The Domainer Accusation

Welcome to the fascinating world of domaining, where website names become lucrative assets. Imagine you’re a collector, each domain you own serving as a unique billboard in the digital landscape, enticing visitors with ads. For just \$10 a year per domain, if any of your domains make an average of \$0.20 per week, it will be making you a profit.

Here’s where the plot thickens. Meet Gabrielle, a shrewd domainer with a portfolio of 1,000 domains. The table below contains 8 weeks of revenue for each of the domains in her collection.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

Bob, a fellow domainer, proposes to buy Gabrielle's domains that are trending to be profitable during week 7. Gabrielle accepts the offer and when about 300 domains end up meeting the criteria, Bob purchases them.

But the following week brings an unexpected twist. The domains purchased by Bob experience a sharp decline in revenue. Bob, feeling duped, accuses Gabrielle of orchestrating a click-fraud operation to artificially boost the week 7 revenue. Gabrielle stands by her innocence, claiming no foul play.

Your Role

Your assignment, as an aspiring data scientist, is to delve into the data to see if it supports Bob's accusation. Is this really a case of deception? Utilize your skills to dissect the data and derive insights that could either validate or debunk Bob's claim. Prepare to present your findings convincingly, for the proof lies within the data.

Your Tools

You'll be using SQL queries to analyze the dataset by running them in the 'Query SQL' pane to the right. If SQL isn't your strongest suit, fear not: ChatGPT is here to help. Provide it with the "create table" statement below and ask your data-related questions and ChatGPT will help generate the necessary SQL statements to retrieve the answers.

This is a unique opportunity to apply your analytical skills to a real-world scenario. Are you ready to unravel the mystery behind this domain dilemma? Let's get to it.

ChatGPT's Attempt at Solving the Challenge

By wording my request carefully, ChatGPT's initial response is stellar, producing functional and helpful SQL code that summarizes the weekly revenue for the domains that were purchased. This is where ChatGPT shines and what scares programmers into thinking their careers are in jeopardy: careful prompting can produce functional code in seconds. Following that are selections from the chat history that, while still impressive, don't quite answer the question. I attempted to prompt it into discovering that the exercise data could have been random (it actually was; there was nothing special about week 7).

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

Given the following data structure, what SQL can I use to retrieve the total revenue by week for domains that made over 0.20 revenue in week 7? CREATE TABLE DomainRev (DomainID INTEGER, Week INTEGER, Revenue REAL);



Data Science Mentor

To retrieve the total revenue by week for domains that made over 0.20 revenue in Week 7, you'll need to perform a couple of SQL operations. The steps involve filtering the domains based on their revenue in Week 7 and then aggregating their revenue across different weeks. Here's an SQL query that achieves this:

```
sql Copy code

SELECT Week, SUM(Revenue) as TotalRevenue
FROM DomainRev
WHERE DomainID IN (
    SELECT DomainID
    FROM DomainRev
    WHERE Week = 7 AND Revenue > 0.20
)
GROUP BY Week;
```

Let's break down this query:

1. **Subquery:** ``SELECT DomainID FROM DomainRev WHERE Week = 7 AND Revenue > 0.20``
 - This subquery selects the ``DomainID``s of all domains that made more than 0.20 in revenue during Week 7.
2. **Main Query:**
 - ``FROM DomainRev WHERE DomainID IN (...)``: The main query uses the list of ``DomainID``s obtained from the subquery to filter the ``DomainRev`` table.
 - ``SELECT Week, SUM(Revenue) as TotalRevenue``: This part aggregates (sums) the revenue for each week.
 - ``GROUP BY Week``: This groups the results by week, so you get a total revenue for



BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

The results show that week 7 revenue is much higher for these domains. Can you think of any reason why that would be?



Data Science Mentor

Observing a significantly higher revenue in Week 7 for certain domains, as indicated by your SQL query results, can be attributed to several potential factors. It's essential to approach this with a scientific mindset, considering multiple hypotheses and evaluating them against the available data. Here are some possibilities:

1. **Seasonal or Cyclical Factors:** There might be seasonal trends or specific events during Week 7 that drive higher traffic or sales, leading to increased revenue. This could include holidays, special shopping days, or industry-specific cycles.
2. **Marketing or Promotional Campaigns:** The domains may have been part of a marketing or promotional campaign during Week 7, leading to increased visibility and, consequently, higher revenue.
3. **Changes in Domain Offerings or Pricing:** The domains might have introduced new products, services, or pricing changes that resonated well with their audience during that week.
4. **External Events:** Events like news coverage, social media trends, or viral content can significantly impact traffic and revenue. If such an event is related to the domain's niche, it could cause a temporary spike in revenue.
5. **Platform or Algorithm Changes:** If these domains rely on search engines or social media platforms for traffic, changes in algorithms or platform policies could have temporarily boosted their visibility and revenue.
6. **Data Anomalies or Errors:** Always consider the possibility of data recording errors, anomalies, or changes in data collection methods. Sometimes what appears as a trend can be a data discrepancy.
7. **Competitor Activity:** The activities of competitors, such as a major competitor experiencing downtime or receiving negative publicity, can inadvertently benefit these domains.



BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

Are there any statistical pitfalls I should watch out for?



Data Science Mentor

When analyzing statistical data, such as your SQL query results, it's important to be aware of common pitfalls that can lead to incorrect conclusions. Here are nine key pitfalls to watch out for:

1. **Using Bad Data:** Ensure that the data you're working with is accurate, relevant, and appropriate for your analysis.
2. **Putting Data Before Theory:** Don't overlook the underlying theories or assumptions that should guide your analysis. Data should be used to test theories, not replace them.
3. **Worshipping Math:** Mathematical sophistication does not guarantee better results. The appropriateness and application of the mathematical methods are what matter.
4. **Worshipping Computers:** Computers are tools that depend on the quality of the input data and the soundness of the algorithms used. They are not infallible.
5. **Torturing Data:** Avoid manipulating or excessively mining the data to find a result that may not be there.
6. **Fooling Yourself:** Be aware of your biases and don't let them lead you to see results that aren't there.
7. **Confusing Correlation with Causation:** Just because two variables are correlated does not mean one causes the other.
8. **Being Surprised by Regression Toward the Mean:** Understand that extreme observations are likely to be followed by more central ones.
9. **Doing Harm:** Consider the ethical implications of your analysis and the potential impact of your conclusions.

These pitfalls serve as important reminders to approach data analysis with caution, skepticism, and a commitment to ethical practices.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

Ah, the 9 Pitfalls of Data Science from the incredible minds of Jay Cordes and Gary Smith! Any guesses as to which of these pitfalls might have affected my results?



Data Science Mentor

Based on your observation of unusually high Week 7 revenue for certain domains, several of the "9 Pitfalls of Data Science" outlined by Jay Cordes and Gary Smith might be relevant. Let's consider which ones could potentially affect your results:

1. **Using Bad Data:** If the data for Week 7 or other weeks is inaccurate or incomplete, this could lead to misleading conclusions. Ensure the data's integrity and completeness.
2. **Putting Data Before Theory:** Before jumping to conclusions based on the data, consider any theoretical explanations. For instance, is there a plausible reason for a revenue spike in Week 7?
3. **Confusing Correlation with Causation:** If there's a correlation between certain domains and high Week 7 revenue, it doesn't necessarily mean one causes the other. There might be underlying factors influencing both.
4. **Being Surprised by Regression Toward the Mean:** If Week 7's revenue is an outlier, subsequent weeks might show a return to average levels. Be cautious about overinterpreting a single week's data as a trend.
5. **Fooling Yourself:** Be mindful of confirmation bias. If you have a preconceived notion about what the data should show (e.g., expecting certain domains to perform better), it can skew your interpretation of the results.

While other pitfalls like "Worshipping Math" or "Worshipping Computers" may not be directly relevant in this context, it's always good to be aware of them. The key is to approach the data with a critical eye, considering multiple hypotheses and being aware of how biases and data quality can influence your analysis.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

Regarding #2, one plausible reason for the spike was given: the domains were for sale and week 7 was the time period used to determine how profitable, and therefore, what their price should be (it's a multiple of the revenue). The seller is suspected of using click-fraud to boost the revenue in order to get a higher selling price. How can I determine if this is the true cause for the revenue spike?



Data Science Mentor

Determining if click-fraud was the true cause for the revenue spike in Week 7 requires a multi-faceted approach, combining data analysis, critical thinking, and possibly external investigation. Here are some steps you can take:

1. **Analyze Traffic Data:** If available, examine the web traffic data for these domains during Week 7. Look for unusual patterns such as a sudden spike in traffic, especially from specific sources or geographic locations, which could indicate fraudulent activity.
2. **Examine User Engagement Metrics:** Beyond just the number of clicks, look at engagement metrics like bounce rate, session duration, and pages per session. If there's a high number of clicks but very low engagement, it could suggest non-genuine traffic.
3. **Compare Historical Data:** Analyze the revenue and traffic patterns for these domains in weeks before and after Week 7. A stark contrast in behavior without any legitimate explanation (like marketing campaigns or viral content) could raise suspicions.
4. **Check for Anomalies in Click Sources:** If the clicks predominantly come from suspicious sources, like known click farms or through automated bots, this could be indicative of fraud.
5. **Correlation with External Events:** Determine if there were any external events during Week 7 that could explain the revenue spike, like promotional activities, advertisements, or news coverage. The at ↓ ce of such events might strengthen the click-fraud hypothesis.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

The historical revenue looks like this: `Week RevenueForPurchasedDomains`

1 0.15

2 0.15

3 0.15

4 0.15

5 0.15

6 0.16

7 0.25

8 0.16



Data Science Mentor

Given the historical revenue data for the purchased domains, there are some observations and analyses you can perform to investigate the possibility of click-fraud or artificial inflation of revenue in Week 7:

1. **Stable Revenue Pattern with a Sudden Spike:** The revenue from Weeks 1 to 6 is quite stable, with a slight increase in Week 6. The jump in Week 7 is notable and deviates significantly from the previous pattern. This sudden spike is a potential red flag, especially in the context of your suspicion regarding click-fraud.
2. **Regression to Previous Levels:** Post Week 7, the revenue regresses back to levels similar to Week 6. This reversion is interesting; if the Week 7 spike was due to genuine factors like increased market interest or successful marketing efforts, you might expect a more gradual decline or sustained higher revenue levels.
3. **Analyzing Week 7 in Detail:**
 - **Traffic Analysis:** If you have access to traffic data for Week 7, check for anomalies like a sudden surge in traffic, unusual traffic sources, or geographic inconsistencies.
 - **Engagement Metrics:** Investigate engagement metrics during Week 7. If the traffic quality was poor (e.g., high bounce rates, very short session durations), it might indicate non-genuine traffic.
4. **Contextual Factors:** Consider what was happening around Week 7. Was there any event, marketing campaign, or external factor  that could explain the spike? If not, the case for click-fraud becomes stronger.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

5. **Click-Fraud Detection Tools:** Utilize specialized tools or services that can detect patterns indicative of click-fraud. These tools can often identify bot traffic or other forms of artificial engagement.
6. **Expert Consultation:** Given the complexities involved in detecting click-fraud, you might consider consulting with a digital fraud expert or a data scientist who specializes in this area.
7. **Seller's Motive:** The motive for inflating Week 7's revenue is clear – to enhance the selling price of the domains. This motive, combined with the unusual revenue pattern, adds weight to the suspicion of click-fraud.

In summary, while the data strongly suggests an anomaly in Week 7, proving click-fraud conclusively requires a combination of data analysis, contextual understanding, and possibly the use of specialized tools. The consistent pattern before and after Week 7, combined with the spike in revenue during the critical valuation period, certainly raises suspicions that merit a thorough investigation.



You

Can you think of any SQL statements that would help prove or disprove any of the pitfalls?



...

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



Data Science Mentor

To investigate the potential pitfalls and the suspicion of click-fraud using SQL, you'll need to delve deeper into the data. Assuming you have access to more detailed data like traffic sources, user engagement metrics, and other relevant information, here are some SQL queries that could be helpful:

1. Analyze Traffic Sources for Week 7:

```
sql Copy code

SELECT TrafficSource, COUNT(*) as Visits
FROM TrafficData
WHERE Week = 7 AND DomainID IN (SELECT DomainID FROM DomainRev WHERE Week = 7)
GROUP BY TrafficSource;
```

This query will help identify where the traffic in Week 7 was coming from, highlighting any unusual sources.

2. Compare Engagement Metrics Across Weeks:

```
sql Copy code

SELECT Week, AVG(SessionDuration) as AvgSessionDuration, AVG(BounceRate) as AvgBounceRate
FROM EngagementData
WHERE DomainID IN (SELECT DomainID FROM DomainRev WHERE Week = 7)
GROUP BY Week;
```

This query averages session duration and bounce rate for each week, allowing you to compare user engagement before, during, and after Week 7.

3. Check for Sudden Increases in New Users or Specific Geographies:

```
sql Copy code

SELECT Week, COUNT(DISTINCT UserID) as NewUsers, Region
FROM UserData
WHERE Week BETWEEN 5 AND 9
```

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

How could I find out if this is just an example of regression towards the mean?



Data Science Mentor

To investigate if the revenue spike in Week 7 is an example of regression toward the mean, you would typically look for a pattern where an extreme measurement (like the high revenue in Week 7) is followed by measurements that are closer to the average. In the context of your data, this would mean checking if the revenue in weeks following Week 7 returns closer to the levels observed before Week 7.

Here are a few steps and a SQL query that can help you analyze this:

1. **Calculate the Mean Revenue Excluding Week 7:**

First, find out the average revenue for weeks excluding Week 7. This gives you a baseline to compare against.

2. **Compare Week 7 Revenue to Other Weeks:**

Compare the Week 7 revenue to the average calculated in step 1. A significant deviation would confirm that Week 7 is an outlier.

3. **Analyze Post-Week 7 Revenue:**

Look at the revenue for weeks after Week 7. If it regresses towards the mean calculated in step 1, it could be a case of regression toward the mean.

Here's a SQL query to assist with these steps:

```
sql Copy code

SELECT Week,
       AVG(Revenue) OVER (ORDER BY Week ROWS BETWEEN UNBOUNDED PRECEDING
                          AVG(Revenue) OVER (ORDER BY Week ROWS BETWEEN UNBOUNDED PRECEDING
Revenue
FROM DomainRev
ORDER BY Week;
```

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

This SQL query returns a row for every domain in the database. Should the data be grouped somehow?



Data Science Mentor

Yes, to effectively analyze the overall trend of revenue across weeks, you should group the data by week. This approach will aggregate the revenue across all domains for each week, allowing you to observe the overall trend rather than individual domain behaviors. Here's how you can modify the SQL query to achieve this:

```
sql Copy code

SELECT Week,
       AVG(Revenue) as AverageRevenue,
       (SELECT AVG(Revenue) FROM DomainRev WHERE Week < 7) as AverageBe
       (SELECT AVG(Revenue) FROM DomainRev WHERE Week > 7) as AverageAf
FROM DomainRev
GROUP BY Week
ORDER BY Week;
```

This revised query will provide:

- **Week**: The week number.
- **AverageRevenue**: The average revenue for that week across all domains.
- **AverageBeforeWeek7**: The average revenue across all domains for weeks before Week 7. This value will be the same for each row and serves as a baseline to compare against Week 7.
- **AverageAfterWeek7**: The average revenue across all domains for weeks after Week 7. This also remains constant across rows and helps to compare the post-Week 7 period.



BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

 You

The results show that revenue is pretty much the same for all domains before and after week 7. If you recall, the spike in revenue we saw before was only for the domains that were profitable during week 7 (revenue > 0.2). So what is this telling us? The result of your query is below...

	Week	AverageRevenue	AverageBeforeWeek7	AverageAfterWeek7
--	------	----------------	--------------------	-------------------

1	0.151080000000000013	0.150349999999999835	0.155250000000000014
2	0.143210000000000014	0.150349999999999835	0.155250000000000014
3	0.15153	0.150349999999999835	0.155250000000000014
4	0.155630000000000001	0.150349999999999835	0.155250000000000014
5	0.151220000000000027	0.150349999999999835	0.155250000000000014
6	0.149430000000000004	0.150349999999999835	0.155250000000000014
7	0.149220000000000024	0.150349999999999835	0.155250000000000014
8	0.155250000000000014	0.150349999999999835	0.155250000000000014



 Data Science Mentor

The results of the query suggest that when considering all domains together, the average revenue before, during, and after Week 7 does not show significant variations. However, this finding contrasts with your initial observation of a revenue spike in Week 7 for domains that had revenue greater than 0.20 in that week. This discrepancy indicates a couple of key points:

1. **Selective Analysis of High-Performing Domains:** The initial observation of a revenue spike was based on a subset of domains (those with revenue > 0.20 in Week 7). This subset might not represent the general trend across all domains. It's a classic case of how analyzing a subset of data can lead to different conclusions than analyzing the entire dataset.
2. **Overall Stability in Revenue:** The results showing similar average revenues across all weeks for all domains suggest that, on average, most domains did not experience significant fluctuations in revenue. This stability might mask the behavior of a smaller subset of domains that did have significant  anges, like the ones you initially focused on.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



Data Science Mentor

The original SQL query you used indeed focuses on a specific subset of domains: those that had a revenue greater than 0.20 in Week 7. The dramatic increase in revenue for this subset in Week 7, as revealed by your query, contrasts with the overall stability observed when considering all domains. This contrast tells us a few things:

1. **Specific Subset Behavior:** The dramatic revenue increase is specific to a subset of domains (those successful in Week 7). This suggests that whatever caused the revenue spike in Week 7 was particularly impactful for these domains, but not necessarily for others.
2. **Potential Influence of External Factors:** The spike could be attributed to external factors impacting only these domains. This might include targeted marketing efforts, events, or other influences that did not affect the broader set of domains.
3. **Highlighting Outliers:** The domains with more than 0.20 revenue in Week 7 might be outliers in the overall dataset. Their performance in Week 7 is not indicative of the general trend across all domains.
4. **Importance of Segment Analysis in Data:** This situation illustrates why it's important to consider different segments of your data separately. Analyzing the entire dataset can sometimes mask significant patterns present in subsets of the data.
5. **Regression Toward the Mean Consideration:** Given that these domains experienced a spike in Week 7 followed by a return to lower revenue levels (assuming Week 8 and onwards follow the general trend), it could be seen as a form of regression toward the mean for this specific subset. However, given the context you've provided (e.g., potential click-fraud for higher selling prices), this might be more indicative of artificial inflation rather than a natural statistical phenomenon.

In conclusion, the behavior of this subset of domains (those successful in Week 7) is distinct from the overall trends in your dataset. This distinction underscores the need to carefully select which segments of data *  analyze, depending on the specific questions or hypotheses you are investigating.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

You mention that context and the incentives to sell for a higher price suggest that regression towards the mean may not be the cause. Is there no direct way to rule out regression toward the mean using the data we have?



Data Science Mentor

To directly rule out regression toward the mean using the data you have, we need to understand the nature of regression toward the mean and then examine if the data patterns fit this phenomenon. Regression toward the mean occurs when an extreme value on one measurement is followed by a value closer to the average on the next measurement, purely due to chance variability.

In your case, you have revenue data for domains across different weeks. To assess for regression toward the mean, you'd typically look for:

1. An extreme value in Week 7 for a specific set of domains (which you've identified).
2. A return to values closer to the average in subsequent weeks for those same domains.

However, the context here is crucial:

- **Context of Week 7 Spike:** If the Week 7 spike is due to external manipulation (like suspected click-fraud), this isn't a random fluctuation but a deliberate action. Regression toward the mean typically applies to natural variations, not those induced by external factors.
- **Pre and Post Week 7 Data:** For regression toward the mean, we'd expect the revenue to return closer to the average values observed in Weeks 1-6 in the weeks following Week 7. However, if this return to "normal" is because the artificial inflation stopped, it's not a regression toward the mean in a statistical sense but rather the cessation of an abnormal condition.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

Given these considerations, you can look at the data post-Week 7:

- If revenue for these specific domains in weeks after Week 7 returns to levels similar to weeks 1-6, it might seem like regression toward the mean. But given the context of potential manipulation in Week 7, this would more likely represent the end of the artificial inflation period.
- If the revenue doesn't return to previous levels, it suggests that other factors (beyond potential click-fraud) might be influencing the revenue patterns.

In conclusion, while you can observe patterns in the data that resemble regression toward the mean, the context of suspected click-fraud implies that these patterns may not be due to random chance but rather to specific actions taken during Week 7. Therefore, it's challenging to attribute the observed pattern solely to regression toward the mean using the available data.



You

How about this: is there a way to exonerate the accused by demonstrating that the spike in revenue for the purchased domains in week 7 is consistent with the null hypothesis that revenue is randomly distributed? In other words, can we show that there is nothing unexpected about the spike in revenue?



Data Science Mentor

To assess whether the spike in revenue for the purchased domains in Week 7 is consistent with the null hypothesis that revenue is randomly distributed (and thus not unexpected), you can perform a statistical significance test. The goal is to determine if the observed spike in Week 7 could reasonably occur as part of natural random variations in the data.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

Given the limitations of your data (only having ``DomainID``, ``Week``, and ``Revenue``), the most straightforward approach would be to use a statistical test like a t-test or an ANOVA, depending on the structure of your data. However, these tests require more granular data, such as individual transaction amounts or the number of transactions, which you don't have.

Here's a simplified approach using the data available:

1. **Calculate Mean and Standard Deviation for Weeks Excluding Week 7:**

Calculate the average revenue and standard deviation for the weeks before and after Week 7 (excluding Week 7).

2. **Compare Week 7 Revenue Against This Baseline:**

Compare the average revenue for Week 7 against this baseline to see if it falls within the expected range of variation.

3. **Perform a Simplified Z-test:**

Assuming the number of data points (domains) is large, you could perform a simplified Z-test to see if the Week 7 revenue is significantly different from other weeks. The Z-score is calculated as $Z = \frac{(\text{MeanWeek 7} - \text{Mean}_{\text{other weeks}})}{\text{StdDev}_{\text{other weeks}} / \sqrt{n}}$, where n is the number of domains.

4. **Interpret the Results:**

- If the Week 7 revenue is within the range of what's expected based on the standard deviation, it might suggest that the spike could be due to chance.
- If the revenue is significantly higher (i.e., the Z-score indicates a statistically significant difference), this suggests the spike might not be due to random variation.

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin



You

Wouldn't statistical tests just show that the spike in week 7 revenue is significantly higher than the other weeks? I'm less interested in "how big is the spike?" than in "are spikes in revenue like this unusual or unexpected?" In other words, if the weekly revenue for the domains was completely random, would we see the same "suspicious" spike?



Data Science Mentor

To address your question, we need to determine if a spike like the one observed in Week 7 could occur as a result of random variation in revenue. In other words, you're asking whether such a spike is within the realm of what might be expected by chance, given the nature of your data. This requires a slightly different statistical approach, focusing more on the probability of observing such a spike under the assumption of randomness.

Here's a way to approach this:

1. **Randomness Assumption:** First, assume that weekly revenue is randomly distributed. This assumption implies that each week's revenue is an independent random event with a certain mean and variance.
2. **Simulation or Bootstrap Method:** Since you have limited data, a good approach might be to simulate many sequences of weekly revenues based on the distribution you observe in your data (excluding the spike week). Alternatively, you could use a bootstrap method, where you repeatedly sample (with replacement) from your existing data to create simulated datasets.
3. **Analyze Simulated Data:** For each simulated sequence, check if there are any spikes similar to or larger than the one observed in Week 7. Count how often such spikes occur.
4. **Calculate Probability:** The proportion of simulations where a spike as large as or larger than the Week 7 spike occurs gives you an estimate of how likely such a spike is under the randomness assumption.



BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

5. Interpret the Results:

- If large spikes occur frequently in your simulations, this suggests that the Week 7 spike could plausibly occur due to random variation.
- If such spikes are rare in the simulations, this suggests that the Week 7 spike is unusual and might not be due to chance.

This approach essentially uses the concept of hypothesis testing, where your null hypothesis is that the revenue data is randomly distributed and spikes like the one observed are not unusual. The result of the simulation or bootstrap analysis will give you evidence for or against this null hypothesis.

Keep in mind, this approach has limitations and assumptions:

- It assumes that the other weeks' revenue data can be used as a proxy for the random distribution of revenue.
- It does not account for external factors that might cause revenue spikes.
- The accuracy of the results depends on the number of simulations and the representativeness of your existing data.

Performing this kind of analysis typically requires the use of statistical software or programming languages like R or Python, which have capabilities for running simulations and complex statistical analyses.

Discussion

The *Regression Towards the Mean* chapter in *The 9 Pitfalls of Data Science* includes this paragraph:

“Clients often thought portfolio sellers were ripping them off because the collection of names they bought experienced a sudden decline in revenue after the sale was finalized. The Eighth Pitfall of Data Science warns buyers to expect that when they buy the best-performing domain names. If, instead, they bought the worst-performing domain names, they would likely be pleasantly surprised when revenue regresses upward after the purchase.”

BECOMING A DATA SCIENTIST IN THE AGE OF AI

Where Bots End, You Begin

Despite having read this and also being prompted repeatedly to consider *Regression Towards the Mean*, ChatGPT didn't come up with reasonable ways to test whether or not the spike in revenue we were analyzing could be explained by randomness until I explicitly asked it to. And even then, it came up with a complicated way that suggested that simulating random data and then measuring the frequency of "suspicious" revenue spikes given various assumptions was the way to go. What I was looking for was something straight-forward like changing the weekly revenue that we're observing to week 6, week 5, etc... to see if the same spike occurs every week (given the same filter on domains getting over \$0.20). Putting this together with the fact that overall revenue was flat for all of the domains means that the hypothesis of click-fraud is very difficult to justify.

If the seller were using click-fraud to artificially boost revenue for a random subset of domains in week 7, the following must also have been true:

1. The seller also boosted revenue for a random subset of domains in weeks 1, 2, 3, 4, 5, 6, and 8. Perhaps to mask the mischievous activity in week 7?
2. The seller also carefully and simultaneously **removed** revenue from another subset of domains such that when all of the revenue is totalled up, there is no suspicious increase in revenue.

If this evidence were presented in a court of law, there would definitely be reasonable doubt in the story that the seller jumped through all of these hoops in order to mask the click-fraud. In order to really seal the case, the defense could also create a random set of data and show that the exact same spike in revenue occurs whenever you select the top performing domains in any given week.

As mentioned in the beginning, the bottom line is that ChatGPT has tons of **knowledge**, but is unable to **think**. In this case, I could lead the ChatGPT horse to the regression towards the mean water, but I couldn't make it drink. In order to avoid fooling ourselves, we need the ability to suppose we're wrong and then come up with ways to try to rule it out. In this case, we needed to look at the data in different ways to find out if there's really something weird going on in week 7 or if it's just a surprising spike that exists in random data when you zoom in on the top performers. There's no simple way to "be sure" that you've got the right idea, so it's unlikely that following automated instructions or reciting memorized knowledge will ever be enough. You have to think.