Data Science Projects

Project 1 : Exploring Hackers News Posts

Overview

In this data science project, you'll analyze a dataset of submissions to Hacker News, a popular technology-focused news aggregator. Using Python and Jupyter Notebook, you'll explore patterns in post creation times, compare engagement levels between different post types, and identify the best times to post for maximum comments. This project will strengthen your skills in data manipulation, analysis, and interpretation, providing valuable experience for aspiring data scientists.

Tools and Technologies

- Python
- Jupyter Notebook

Prerequisites

- String manipulation and basic text processing
- Working with dates and times using the datetime module
- Using loops to iterate through data collections
- Basic data analysis techniques like calculating averages and sorting
- Creating and manipulating lists and dictionaries

Step-by-Step Instructions

1. Load and explore the Hacker News dataset, focusing on post titles and creation times
2. Separate and analyze 'Ask HN' and 'Show HN' posts
3. Calculate and compare the average number of comments for different post types
4. Determine the relationship between post creation time and comment activity
5. Identify the optimal times to post for maximum engagement

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Manipulating strings and datetime objects in Python for data analysis
- Calculating and interpreting averages to compare dataset subgroups
- Identifying time-based patterns in user engagement data
- Translating data insights into practical posting strategies

Data Set Link:

https://www.kaggle.com/datasets/hacker-news/hacker-news-posts

Project 2: Exploring e-Bay Car Sales Data

Overview

In this data science project, you'll analyze a dataset of used car listings from eBay auto dataset, a classifieds section of the German eBay website. Using Python and pandas, you'll clean the data, explore the included listings, and uncover insights about used car prices, popular brands, and the relationships between various car attributes. This project will strengthen your data cleaning and exploratory data analysis skills, providing valuable experience in working with real-world, messy datasets.

Tools and Technologies

- Python
- Jupyter Notebook
- NumPy
- pandas

Prerequisites

- Loading and inspecting data using pandas
- Cleaning column names and handling missing data
- Using pandas to filter, sort, and aggregate data
- Creating basic visualizations with pandas
- Handling data type conversions in pandas

Step-by-Step Instructions

1. Load the dataset and perform initial data exploration
2. Clean column names and convert data types as necessary
3. Analyze the distribution of car prices and registration years
4. Explore relationships between brand, price, and vehicle type
5. Investigate the impact of car age on pricing

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Cleaning and preparing a real-world dataset using pandas
- Performing exploratory data analysis on a large dataset
- Creating data visualizations to communicate findings effectively
- Deriving actionable insights from used car market data

Data Set Link:

https://drive.google.com/file/d/1A46CjH7DrZvyS0Un-8BvlmrAUVt0elM-/view?usp=sharing

Project 3: Finding Heavy Traffic Indicators on I-94

Overview

In this data science project, you'll analyze a dataset of westbound traffic on the I-94 Interstate highway between Minneapolis and St. Paul, Minnesota. Using Python and popular data visualization libraries, you'll explore traffic volume patterns to identify indicators of heavy traffic. You'll investigate how factors such as time of day, day of the week, weather conditions, and holidays impact traffic volume. This project will enhance your skills in exploratory data analysis and data visualization, providing valuable experience in deriving actionable insights from real-world time series data.

Tools and Technologies

- Python
- Jupyter Notebook
- pandas
- Matplotlib
- seaborn

Prerequisites

- Data manipulation and analysis using pandas
- Creating various plot types (line, bar, scatter) with Matplotlib
- Enhancing visualizations using seaborn
- Interpreting time series data and identifying patterns
- Basic statistical concepts like correlation and distribution

Step-by-Step Instructions

1. Load and perform initial exploration of the I-94 traffic dataset
2. Visualize traffic volume patterns over time using line plots
3. Analyze traffic volume distribution by day of the week and time of day
4. Investigate the relationship between weather conditions and traffic volume
5. Identify and visualize other factors correlated with heavy traffic

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Creating and interpreting complex data visualizations using Matplotlib and seaborn
- Analyzing time series data to uncover temporal patterns and trends
- Using visual exploration techniques to identify correlations in multivariate data
- Communicating data insights effectively through clear, informative plots

Data Set Link:

https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume

Project 4: Story Telling Data Visualization on Exchange Rates

Overview

In this data science project, you'll create a storytelling data visualization about Euro exchange rates against the US Dollar. Using Python and Matplotlib, you'll analyze historical exchange rate data from 1999 to 2021, identifying key trends and events that have shaped the Euro-Dollar relationship. You'll apply data visualization principles to clean data, develop a narrative around exchange rate fluctuations, and create an engaging and informative visual story. This project will strengthen your ability to communicate complex financial data insights effectively through visual storytelling.

Tools and Technologies

- Python
- Jupyter Notebook
- pandas
- Matplotlib

Prerequisites

- Data manipulation and analysis using pandas
- Creating and customizing plots with Matplotlib
- Applying design principles to enhance data visualizations
- Working with time series data in Python
- Basic understanding of exchange rates and economic indicators

Step-by-Step Instructions

1. Load and explore the Euro-Dollar exchange rate dataset
2. Clean the data and calculate rolling averages to smooth out fluctuations
3. Identify significant trends and events in the exchange rate history
4. Develop a narrative that explains key patterns in the data
5. Create a polished line plot that tells your exchange rate story

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Crafting a compelling narrative around complex financial data
- Designing clear, informative visualizations that support your story
- Using Matplotlib to create publication-quality line plots with annotations
- Applying color theory and typography to enhance visual communication

Data Set Link:

https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-usd.en.html

Project 5: Clean and Analyze Employee Exit Surveys

Overview

In this data science project, you'll analyze employee exit surveys from the Department of Education, Training and Employment (DETE) and the Technical and Further Education (TAFE) institute in Queensland, Australia. Using Python and pandas, you'll clean messy data, combine datasets, and uncover insights into resignation patterns. You'll investigate factors such as years of service, age groups, and job dissatisfaction to understand why employees leave. This project offers hands-on experience in data cleaning and exploratory analysis, essential skills for aspiring data analysts.

Tools and Technologies

- Python
- Jupyter Notebook
- pandas

Prerequisites

- Basic pandas operations for data manipulation
- Handling missing data and data type conversions
- Merging and concatenating DataFrames
- Using string methods in pandas for text data cleaning
- Basic data analysis and aggregation techniques

Step-by-Step Instructions

1. Load and explore the DETE and TAFE exit survey datasets
2. Clean column names and handle missing values in both datasets
3. Standardize and combine the "resignation reasons" columns
4. Merge the DETE and TAFE datasets for unified analysis
5. Analyze resignation reasons and their correlation with employee characteristics

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Applying data cleaning techniques to prepare messy, real-world datasets
- Combining data from multiple sources using pandas merge and concatenate functions
- Creating new categories from existing data to facilitate analysis
- Conducting exploratory data analysis to uncover trends in employee resignations

Data Set Link:

https://data.gov.au/dataset/ds-qld-fe96ff30-d157-4a81-851d-215f2a0fe26d/details?q=

Project 6: Star Wars Survey

Overview

In this beginner-level data science project, you'll analyze survey data about the Star Wars film franchise. Using Python and pandas, you'll clean and explore data collected by FiveThirtyEight to uncover insights about fans' favorite characters, film rankings, and how opinions vary across different demographic groups. You'll practice essential data cleaning techniques like handling missing values and converting data types, while also conducting basic statistical analysis to reveal trends in Star Wars fandom.

Tools and Technologies

- Python
- Jupyter Notebook
- pandas

Prerequisites

- Loading and inspecting data using pandas
- Cleaning column names and handling missing data
- Converting data types in pandas DataFrames
- Filtering and sorting data
- Basic data aggregation and analysis techniques

Step-by-Step Instructions

1. Load the Star Wars survey data and explore its structure
2. Clean column names and convert data types as necessary
3. Analyze the rankings of Star Wars films among respondents
4. Explore viewership and character popularity across different demographics
5. Investigate the relationship between fan characteristics and their opinions

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Applying data cleaning techniques to prepare survey data for analysis
- Using pandas to explore and manipulate structured data
- Performing basic statistical analysis on categorical and numerical data
- Interpreting survey results to draw meaningful conclusions about fan preferences

Data Set Link:

https://drive.google.com/file/d/1s9rFPp_6UGjF5fMccmO4DaxbQNAMN3FW/view?usp=sharing

Project 7: Finding the Best Markets to Advertise In

Overview

In this data science project, you'll analyze survey data from freeCodeCamp to determine the best markets for an e-learning company to advertise its programming courses. Using Python and pandas, you'll explore the demographics of new coders, their locations, and their willingness to pay for courses. You'll clean the data, handle outliers, and use frequency analysis to identify countries with the most potential customers. By the end, you'll provide data-driven recommendations on where the company should focus its advertising efforts to maximize its return on investment.

Tools and Technologies

- Python
- Jupyter Notebook
- pandas

Prerequisites

- Loading and inspecting data using pandas
- Filtering and sorting DataFrames
- Handling missing data and outliers
- Calculating summary statistics (mean, median, mode)
- Creating and manipulating new columns based on existing data

Step-by-Step Instructions

1. Load the freeCodeCamp 2017 New Coder Survey data
2. Identify and handle missing values in the dataset
3. Analyze the distribution of participants across different countries
4. Calculate the average amount students are willing to pay for courses by country
5. Identify and handle outliers in the monthly spending data
6. Determine the top countries based on number of potential customers and their spending power

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Cleaning and preprocessing survey data for analysis using pandas
- Applying frequency analysis to identify key markets
- Handling outliers to ensure accurate calculations of spending potential
- Combining multiple factors to make data-driven business recommendations

Data Set Link:

https://drive.google.com/file/d/1qxvX1SdbKJWpMoUaGYAVf364xovGDWmV/view?usp=sharing

Project 8: Building a Spam Filter using Naive Bayes

Overview

In this data science project, you'll build a spam filter using the multinomial Naive Bayes algorithm. Working with the SMS Spam Collection dataset, you'll implement the algorithm from scratch to classify messages as spam or ham (non-spam). You'll calculate word frequencies, prior probabilities, and conditional probabilities to make predictions. This project will deepen your understanding of probabilistic machine learning algorithms, text classification, and the practical application of Bayesian methods in natural language processing.

Tools and Technologies

- Python
- Jupyter Notebook
- pandas

Prerequisites

- Python programming, including working with dictionaries and lists
- Understand probability concepts like conditional probability and Bayes' theorem
- Text processing techniques (tokenization, lowercasing)
- Pandas for data manipulation
- Understanding of the Naive Bayes algorithm and its assumptions

Step-by-Step Instructions

1. Load and explore the SMS Spam Collection dataset
2. Preprocess the text data by tokenizing and cleaning the messages
3. Calculate the prior probabilities for spam and ham messages
4. Compute word frequencies and conditional probabilities
5. Implement the Naive Bayes algorithm to classify messages
6. Test the model and evaluate its accuracy on unseen data

Expected Outcomes

Upon completing this project, you'll have gained valuable skills and experience, including:

- Implementing the multinomial Naive Bayes algorithm from scratch
- Applying Bayesian probability calculations in a real-world context
- Preprocessing text data for machine learning applications
- Evaluating a text classification model's performance

Data Set Link:

https://archive.ics.uci.edu/dataset/228/sms+spam+collection

Machine Learning Projects

Project 1: Home Value Prediction

Consider a situation where you want to buy or sell a house or are moving to a new city and want to rent a home, but you need to know where to start. Sometimes, you know where to start but must check the source's credibility. Some people from Microsoft also felt the need to create a reliable place to provide all this information online, and "Zillow" was born in 2006. Zillow introduced a "Zestimate" feature a few years later, completely changing the market. Zestimate is a tool that provides the house's worth based on various attributes like public and sales data. Zestimate has information on more than 97 million homes snd as per Zillow, Zestimates are within the range of 10% of the selling price of homes.

Project Idea: In this **Machine Learning project for students**, you will use the Zillows Economics data set to build a house price prediction model with XGBoost based on factors like average income, crime rate, number of hospitals, number of schools, etc. Having completed this top ML project, one should be able to answer questions like top states with the highest rent values, where you should buy/rent a house, Zestimate per square feet, the median rental price for all homes, etc.

Industry: Real Estate

Data Set Link:

https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Project 2: Iris Flowers Classification

Here is one of the most simple machine learning projects, with Iris Flowers being the most straightforward machine learning dataset in the classification problems literature. This machine learning problem is often called the "Hello World" of machine learning. The dataset has numeric attributes, and ML beginners need to figure out how to load and handle data. The Iris dataset is small, easily fits into memory, and does not require any extraordinary transformations or scaling.

Project Idea: The Iris Dataset can be downloaded from the UCI ML Repository—Download Iris Flowers Dataset. The goal of this data science project for beginners is to classify the flowers into three species—Virginia, setosa, or versicolor—based on the length and width of the petals and sepals. By implementing advanced algorithms, you can also add this project to your deep learning projects portfolio.
Industry: Medicine

Data Set Link:

https://www.kaggle.com/datasets/uciml/iris

Project 3: Wine Quality Prediction

It’s known that the older the wine, the better the taste. However, several factors other than age go into wine quality certification, which includes physiochemical tests like alcohol quantity, fixed acidity, volatile acidity, determination of density, pH, and more.

Project Idea: The main goal of this machine-learning project is to build a machine-learning model to predict the quality of wines by exploring their various chemical properties. The wine quality dataset consists of 4898 observations with 11 independent and one dependent variable. After using data visualization techniques, figure out the feature variable space that will serve as an input to the machine learning model. Then, prepare the report and fine-tune the model's hyperparameters to enhance the accuracy.

Industry: Viticulture

Data set Link:

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

Project 4: Loan Eligibility Prediction

Loans are what make the world go round. They are the core business for banks since their main profit comes from interest on loans. Sometimes, to be able to take risks of this sort and sometimes, even to have some worldly pleasures, it becomes necessary for one to apply for a loan. Banks usually have a rigorous process to follow before a loan can be approved.And

they can leverage machine learning methods to predict the eligibility for a loan that someone applies for so that there can be better planning beyond the loan being approved or rejected.

Project Idea: The model for determining loan eligibility prediction has to be trained using a dataset that consists of data including data such as sex, marital status, number of dependents, income, qualifications, credit card history and loan amount to name a few. This project will require training and testing the data model using the method of cross validation. After using data visualization techniques, clean the data and fill in the missing values. This project is an excellent means to learn how to build statistical models such as Gradient Boosting and XGBoost, and also to understand metrics such as ROC Curve, MCC scorer and the like.

Industry: Financial Services

Data Set Link:

https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset

Project 5: Inventory Demand Forecasting

Preparing sufficient inventory is a task that not only restaurants registered on Zomato have to complete. Most companies that offer products have to ensure they have enough to satisfy all their customers. It is essential to have a rough estimate of how much preparation would be enough. This estimation can be achieved by what we call demand forecasting. A demand forecast is vital for all business decisions: sales, finance, production management, logistics, and marketing. If these forecasts are correctly predicted, they can help businesses grow

significantly by allowing them to reach customers with the right products at the right time. It can also help companies in avoiding unnecessary wastage of their resources.

Project Idea: By applying relevant algorithms such as Bagging, Boosting, XGBoost, Support Vector Machines, and more, businesses can make accurate predictions about customer demand. This can significantly improve their inventory management and overall operations.

Industry: Multiple

Data Set Link:

https://drive.google.com/file/d/1_NhXDd-XHkcZjv3JMmo21gD1rvXdZ84Z/view?usp=sharing

Project 6: Credit Card Default Prediction

This is one of the top machine learning projects that aims to predict customers who will default on a loan. Banks may experience loss on credit card products from various sources, and one possible reason for the loss is when customers default on their debt, preventing banks from collecting payments for the services rendered.

Project Idea: In this machine learning project, you will examine a slice of the customer database to determine how many customers will be delinquent in making payments in the next two years. There are various machine learning models for predicting which customers default on a loan so the banks can cancel credit lines for risky customers or decrease the credit limit on the card to minimize losses. These models will also help banks screen which customers can be approved for a credit card.

Industry: Financial Services

Data Set Link:

https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

Project 7: Human Activity Recognition

The smartphone dataset consists of fitness activity recordings of 30 people captured through smartphone-enabled inertial sensors.

Project Idea: This project on machine learning aims to build a classification model that can precisely identify human fitness activities. Working on this machine learning project will help you understand how to solve multi-classification problems.

Industry: Medicine

Data Set Link:

https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones

NLP Projects

Project 1: E-Commerce Product Review (Sentiment Analysis)

This is one of the most popular NLP projects that you will find in the bucket of almost every NLP Research Engineer. The reason for the popularity of a sentiment analysis project is that companies widely use it to monitor the efficacy of their product through customer feedback.

Method: The first step to start designing the Sentiment Analysis system would involve performing EDA over textual data. After that, you will have to use text data processing methods to extract relevant information from the data and remove gibberish. The next step would be to use significant words in the reviews to analyze the sentiment of the reviewer.

Data Set Link:

https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

Project 2: Topic Modeling or Topic Identification

This is a very basic NLP Project that expects you to use NLP algorithms to understand them in depth. The task is to have a document and use relevant algorithms to label the document with an appropriate topic. A good application of this NLP project in the real world is using this NLP project to label customer reviews. The companies can then use the topics of the customer reviews to understand where the improvements should be done on priority.

Method: This project will introduce you to methods of handling textual data and using regex. You will understand how to convert textual data into vectors through methods like TF-IDF and Count vectorizer. You will also learn how to use unsupervised machine learning algorithms like K-Means clustering for grouping similar reviews together.

Data Set Link:

https://www.kaggle.com/code/canggih/topic-modeling?select=voted-kaggle-dataset.csv

Project 3: Disease Diagnosis

If you are looking for NLP in healthcare projects, then this project is a must try. Natural Language Processing (NLP) can be used for diagnosing diseases by analyzing the symptoms and medical history of patients expressed in natural language text. NLP techniques can help in identifying the most relevant symptoms and their severity, as well as potential risk factors and comorbidities that might be indicative of certain diseases.

Method: NLP techniques can be used to extract information from unstructured clinical notes and electronic health records, which can be used to predict and diagnose diseases. This information includes patient demographics, medical history, medication and treatment plans, and laboratory results. You can use NLP to identify specific patterns or signals within the text data that might be indicative of a particular disease or condition.

Data Set Link:

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Bonus Project

Description:

Webscrap the post or tweets from twitter or facebook atleast 1000 tweets or posts to be webscrapped. Apply NLP Techniques like tokonization, removing stop words, Identifying Parts of speech, Context Awareness of words to determine the sentiment of each sentence.

Note: Additional marks will be provided if you do the analysis on regional languages like telugu