Presentation Notes:

Tri:
-So, basically, by processing a historical data set,
we look forward to using supervised machine learning model to predict sales accurately

our null hypothesis is that can we make sales prediction based on features

The reason why we selected this topic:

Due to the business challenges, an automatic prediction will not only help the businesses gain more profits but also the customers to buy products that satisfy them.

This is the historical data that covers sales from 2010-02-05 to 2012-11-01. It has all the information of an ideal ML algorithm
in predict demand should have.

In total, there are 8 different fields, which is listed below:
1. Store - the store number
2. Date - the week of sales
3. Weekly_Sales - sales for the given store
4. Holiday_Flag - whether the week is a special holiday week: 1 - Holiday week; 0 - Non-Holiday week
5. Temperature - Temperature on the day of sale
6. Fuel_Price - Cost of fuel in the region
7. CPI - Prevailing consumer price index
8. Unemployment - The prevailing unemployment rate

**Question want to answer:
Can a supervised machine learning model predict unforeseen sales for the business?

Vick:

- **Walmart Sales Dashboard** shows Total Sales for 45 WMT stores and average of the four features from the Kaggle website.

a. The dashboard also shows seasonality with ebbs and flows over time. Notice the sales increase towards peak summer Jun-Jul and Nov-Dec time frame. We are trying to predict sales for peak season here week 50.
b. Filter is by store so that you can see sales and features by store as well. The filter on the right helps us define results at store level. Also, our machine learning model will forecast at the each of the store level

- **Database** - For us connectivity of dataset is an important element. We took four step process to connect, store, and share our information with the help of databases:
    a. Step 1: Created an AWS RDS online database
    b. Step 2: Created a AWS RDS server on PostgreSQL for connection and a Walmart Sales database
    c. Step 3: Created a database ERD diagram, developed SQL create table and joins table in Visual Studio, and third created actual schema (tables) in PostgreSQL
    d. Step 4: Fed in data from the ETL - Jupyter notebook file into PostgreSQL database. Once data was in PgAdmin4 it fed automatically to AWS RDS, which fed back into Jupyter notebooks with the help of SQLAlchemy.

Frances:

- Why we chose each model: **Linear Regression**- Common model used for forecasting because it can measure the relationship between variables and forecasting. IT can determine trends in the data. Plots a straight line through the prices so as to minimize the distance between prices. **Random Forest**- We chose bc of ability to take in other features(fuel price, unemployment, ect.)
- Metrics used to measure model:
    - RMSE: Standard deviation of the prediction errors. Difference between predicted and actual values. RSME shows how concentrated our data is around the line of best fit. Larger errors have a greater impact on the overall error, lower scores are better (less error)
    - MAE: Measures the average between the absolute differences between the prediction and the actual values . ranges from 0-infinity, lower values better
    - R2: represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. It is measured from 0%-100%. 0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. The higher the R-squared, the better the model fits your data.

    - Naive RMSE:Same as RMSE with the assumption that predictions will mirror previous week sales. This score was introduced to ensure that our model performed up to naive RMSE standard. The lower the RMSE of the model compared to naive RMSE, the better the performance.
- Briefly explain the tools, tech, languages we used
- Takeaways: our hypothesis is correct: we can predict sales for 45 Walmart stores (using Random Forest), we improved our model for RF by decreasing our features (removed the lags and sales difference, used previous weekly sales and the original features) and we increased our estimates (100 to 250 produces about 2% improvement)
- Improvements: use new data with higher correlations (our highest correlation was unemployment with about -.1% but the rest were around .03 and lower), include location to better map the store sales (we can determine population density, avg. income, ect. from locations), use neural networks to improve accuracy (ARIMA, LSTM)

Archana:

- First we used Linear Regression Model and then Random Forest Regression Model on all the sores sales data. The scores of the Random Forest Regression Model were far better than the Linear Regression Model.
- Then we used “lag” on store-1 data only as an initial learning exercise and found out that the results were improved a lot.
- The features are arranged in the order of their importance and it can be seen that the individual weeks and lags play a crucial role in the stores sales forecasting.
- We have chosen three Scenarios:
  - Scenario-1 (Using all input features):
    Avg. R2=31.36%
  - Scenario-2 (Using only Previous Week Sales and Weekly Sales):
    Avg. R2=45.58%
  - Scenario-3(Using Previous Week Sales and original features i.e. unemployment, CPI, temperature, fuel price, holiday):
    Avg. R2=48.05%
- Dashboard shows all the three scenarios results and the week-50th sale by store in descending order.
- Our Random Forest Regression model is able to predict the week-50th sale by using the Naïve forecasting technique with an average R2 equal to 48%. The R2 value can be further improved by increasing the number of estimators.