IIntroductions: SP
Database: Inez
ML: dana
Visualization: Ivan

Greetings and intro:
Good evening class, my name is Shohesh, here're our team members Inez, Dana, and Ivan. Inez will be presenting the data ETL and database, Dana will be presenting the machine learning portion, and Ivan will be presenting the visualization.
Here we go, our topic is Toronto bike demand prediction based on weather conditions. We want to explore bonding between bike demand and weather conditions.

Database :

Slide 5
Good evening, i will explain about the etl process, database and connection to machine learning.

Slide 6
The first one is data extraction, for the bike data we downloaded it from open source portal CKAN. The data format is csv and it's separated to few csv by years.
Like we can see here, the csv consists of bike -sharing user trip data

Slide 7
For the purpose of machine learning, We cleaned the data using jupyter notebook and pandas, (click) The first we need to do is to combine all of the csv to one dataframe , then make sure there is no missing values, after that we group the data. (click)Here is the data that has been cleaned and grouped.

Slide 8
For the weather data we have two data sources, the first one we scraped it from wunderground website, and the second one was downloaded from Meteostat website.

Slide 9

To clean the data we use jupyter notebook and pandas . Here we can see what we do to clean and merge the data. (click)And here is the result.
Slide 10

After we clean the data , we transfer the data to the database using jupyter notebook and pandas. We separate the data into two tables , tbl_bike_data and tbl_weather_data. Here is the printscreen of our table in PostgreSQL. To make sure each team member has the same

database,each team member will run the script that we have made before on their own computer.

Slide 11

Next is our ERD Diagram. Here we can see in more detail what field we have in each table and the relationship between two tables. To join between the two tables we will use date as the key.

Slide 12

For the connection to machine learning , we will connect to the database using SQLAlchemy. Here we use sql query to join the two tables.

I think that's all for the database. The machine learning part will be explained by Dana.

Slide 13

Machine Learning:

First step is to:

- extract data table from database and read it into a dataframe,
- then double check whether the data frame has duplicates or null values, making sure it is cleaned and ready to use in next steps.

Next it's to select features: There are three variables that can represent bike demand, or targets: Trip duration, Counts of bikes, and Counts of trips.
After preliminary processing, we decided to use counts of bikes as the target, y, and all the average weather conditions as independent features, or, X .
Next step is to split the data set into training and testing sets using sklearn's train_test_split library, and standardize the feature sets using StandardScaler library
After that, model training process begins:

- First model to try is the Linear regression model, since it is relatively fast to compute and easy to detect linear relationships. yet it is sensitive to outliers and difficult to capture complex relationships. So to detect any non-linear relationships, we use the decision tree models.

After fitting the model, we evaluate the model performance using metrics including R2 scores, model scores and a few others.
As shown here, Random forest has the highest prediction model score, and relatively high R2 score, Additionally, comparing the mean absolute error, mean squared error, and root mean squared errors, so far the random forest model has relatively lower digits in these metrics, and thus it's a better representation of the dataset.
For future analysis, there are two recommendations:

1: To improve prediction accuracy, these are the recommended models to try-out: BaggingRegressor, ExtraTreesRegressor, and AdaBoostRegressor.

2: To add more weather feature variables into the dataset

3: Eventually come up with a prediction equation, which would take in weather conditions and output estimated demand of bikes.

That's all for the machine learning portion.
Next let's welcome Ivan to show us the Visualizations.(2'30")

Visualization

Hi Guys,

It’s Ivan here.

The single most important thought I want to share with you today is team work. Without our effort to put everything together we couldn’t make it, and it was differential in our group.

Our website was created to support our hard work.

I ‘am going to show you our website right now, and as you can see, our goal was exploring possible correlation between different types of weather conditions, and the demand on bike sharing.

We have included in our website, a great tool to show the result of some questions that we would like to answer. We used Tableau to better show this.

On the top of the Data visualization, we see a Bike Temperature line graphic. Depends on the weather we have the most bike demand among the weather conditions, special in the third quarter of the year.

In our first bar chart we found out which are the main stations for renting the bike, giving us the latitude, longitude and Station name.

So, we discovered the difference between members and casual members, which reflects a lot in the company's revenue. The casual member has the potential to become an annual member, so we can use this information to improve customer service.

We also see growth in the use of the bike share service year after year, which relates to people's concern for a healthier life.

The other graph shows the average trip duration by month, and it’s evident that the months with the highest demand are from May to September, when the temperature is higher.

Folks, that’s what we wanted to share with you today and feel free for any questions.