Amusement Park

- ## Introduction

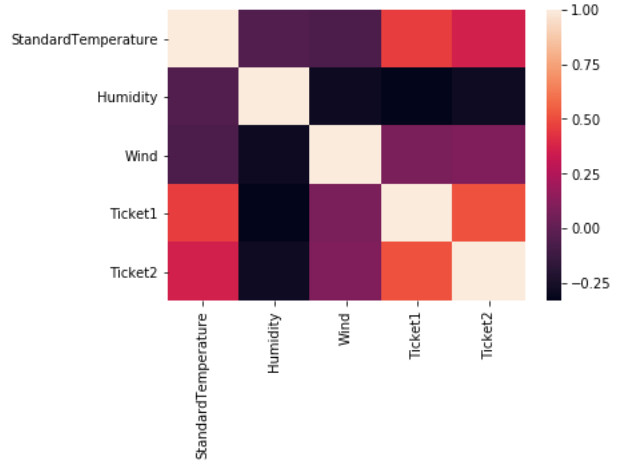
Amusement parks have registered and the unregistered customers who usually visits during definite hours and are also specific about the day on which they paid the visit. Amusement park business have provided us the data regarding the number of visits paid by number of customers who are registered and also who are not even registered and are random visitors. Ticket1 are the count of registered users and Ticket2 is the number of random users who are not registered.
We need to predict the number of tickets sold for two types of tickets – Ticket1 and Ticket2 on the basis of below features:

- Temperature – Standard temperature on that day/time
- Humidity – Humidity experienced on that day/time
- Wind- Speed of wind on that day/time
- Timestamp- Time on which the ticket counts were observed.

- ## Data

We had in total 14000 records in the file that included above features and output variables which include sales of Ticket1 and Ticket2.

During the Data exploration we saw that the features were not correlated. We have the heat map which suggests the same



Whole dataset was divided by using SkLearn library with train test split function, dividing the data set into train test data set. On applying this function, we receive the following output:

Train_x = Data set of 11200 rows with 3 features used for training the model

Train_y = Data set of 11200 rows for response variable 'Ticket1' and 'Ticket2' were used during training process

Valid_x = The dataset of 2800 rows were used as an input variables/features for the validation process

Valid_y= Data set of 2800 rows of actual values which are compared with the predicted values and used for calculating RMSE values.

- **Methodology**
  - ARIMA

    We first tried the approach of time series analysis using ARIMA (AutoRegressive Integrated Moving Average) model. Below are the steps that were followed:
    1. Sort the data by the timestamp feature
    2. Split the data into 80% of training set and 20% of validation set
    3. Train ARIMA model to forecast Ticket1 and Ticket2

    This approach provided us a decent root mean squared error. But the drawback was that we could not use the features other than timestamp. Since the data had four features, we decided to pursue a different model that will consider Temperature, Humidity and Wind as well.

  - Regression

    We tried fitting the features and output variables using linear regression
    Below are the steps that we followed:
    1. Check the correlation between each of the features and the output variables
    2. Split the data into 80% of training set and 20% of validation set
    3. Train the linear regression model and check the root mean squared error
    4. Also checking the R^2 value which will suggest the explaining ability of models through its features.
    5. **Root mean squared error with validation set: 89.34**

    Below is the output we got for the test dataset that was provided:

  - Polynomial Regression

    We tried fitting the features and output variables using Polynomial regression
    Below are the steps that we followed:
    1. Check the correlation between each of the features and the output variables

2. Split the data into 80% of training set and 20% of validation set
3. Train the polynomial regression model and check the root mean squared error
4. Also checking the R^2 value which will suggest the explaining ability of models through its features.
5. We ran the model using different degree. Degree 3 had the best RMSE value compared to the other degrees.
6. **Root mean squared error with validation set: 87.15**

	Ticket1	Ticket2
0	33.0	132.0
1	20.0	107.0
2	23.0	116.0
3	50.0	180.0
4	37.0	159.0
5	13.0	78.0
6	13.0	90.0
7	86.0	260.0
8	26.0	126.0
9	27.0	130.0
10	13.0	97.0