BoomBikes Assignment Subjective Questions

Qn: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

Qn: Why is it important to use drop_first=True during dummy variable creation?

Ans:
This essentially helps in reducing one column out of the total categories in a categorical variable. In case of the assignment, the weekday column has categorical values from 0 - 6. When dummies are created, when drop_first = true is specified, the number of columns created are 6, the values are weekday_1, weekday_2, weekday_3, weekday_4, weekday_5, weekday_6. Anything other than these 6 will be treated as weekday_0. Note that instead of 0 - 6, it is mapped to day value in the implementation in notebook, jfyi).

Technically, as per the learning so far, when there are n levels of categorical values, then n-1 dummies are required to represent those. Internally, since the number of columns are reduced, the correlations created amount the dummy variables are reduced.

However, as far as the reading goes, removing extra columns helps when the level of categorical values is less, but it becomes less clear when there are a few hundred levels.

Qn: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Qn: How did you validate the assumptions of Linear Regression after building the model on the training set?

Qn: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

General Subjective Questions

Qn: Explain the linear regression algorithm in detail.
Ans: In simple terms, linear regression model is the function which can plot a best fit linear line between dependent and independent variables. The linear regression algorithm can only be used when there is a linear relationship between dependent and independent variables.

There are two types of linear regression, they are mainly the below ones

- Simple Linear regression : Where only one independent variable is present and the relationship between this and the dependent variable has to be found out.

- Multiple Linear regression : Where multiple independent variables are present and the model has to find the relationship with the dependent variable.

The equation for the simple linear regression is y = b0 + b1*x + e
b0, b1 are the beta coefficients / parameters. b0 is in fact the intercept of the regression line on y. b1 is slope of the regression line. This equation is modeled from the y = mx + c (line equation) . e is called error term, but normal cases, the desired value for mean of error is 0 and hence the y = b0 + b1*x

In case of multiple linear regression, the equation becomes,
Y = b0 + b1x1 + b2x2 + .. + bnxn

b0, b1 … bn are coefficients of independent variables.

Now there are multiple important assumptions for Linear regression.

Linearity : the dependent and independent variables should be linearly related. Usually a correlation chart or scatter plot can be done to find out this relationship

Normality: the variables should be normally distributed. Q-Q plots can be used to plot this relationship.

Homoscedasticity: Error terms variance should be constant. When the error terms are plotted, should not show a specific pattern such as a funnel shaped or sinusoidal etc.

No multicollinearity : should be present between the independent variables. If it exists, then preferably only the most relevant ones can be included while building the model function.

Usually, hypothesis testing is done on the beta coeff b1 to find out whether the fitted line is significant or not. Need to start by saying b1 is not significant i.e Null hypothesis will be H0: b1 = 0 , then Ha (b1) != 0

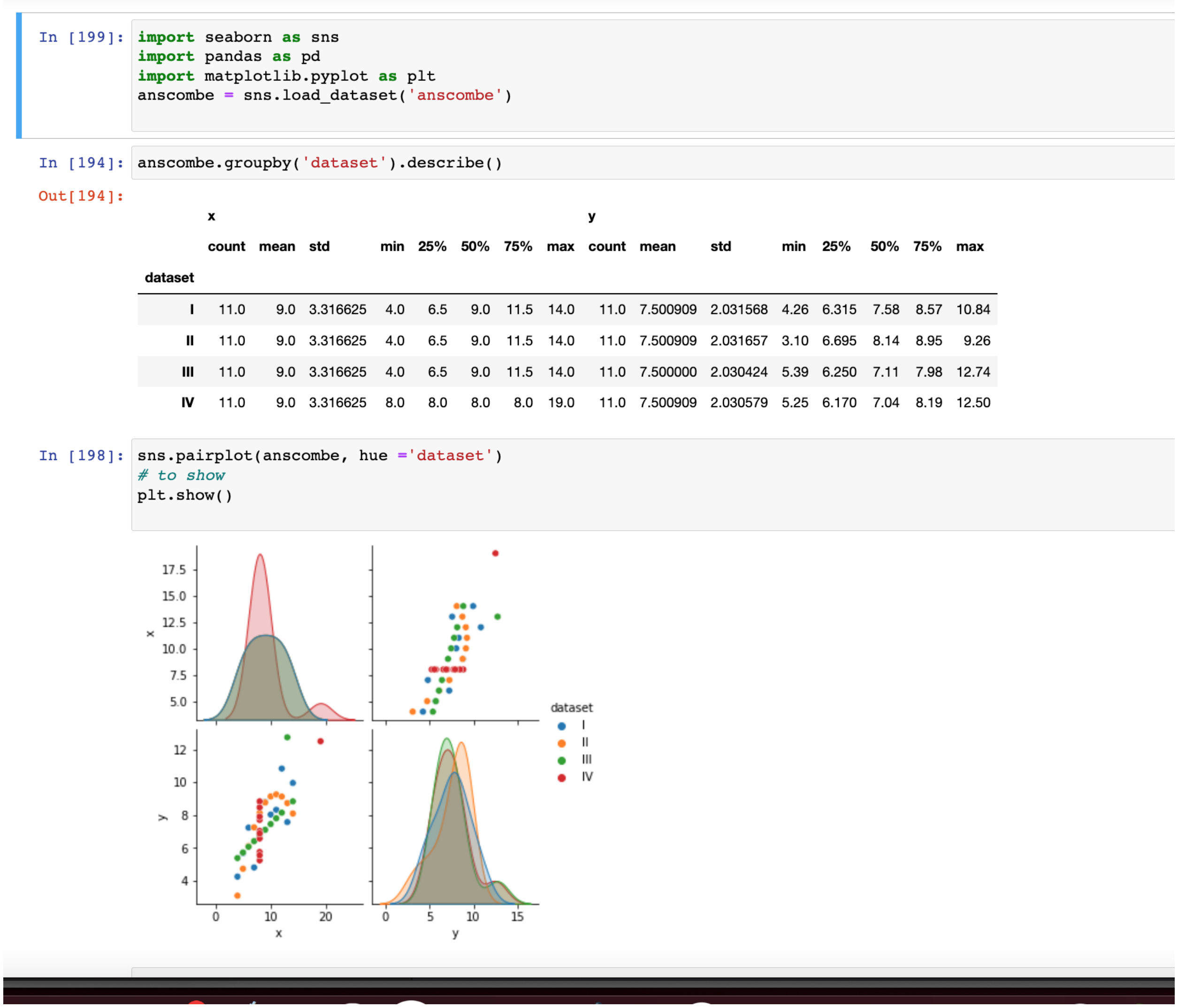
To perform hypothesis test, the p-value to be found for the given beta coeff. Which mainly involves finding the t-score for the mean point 0 (as per the null hypothesis), calculate the p-value from the cumulative probability for the given t-value (using t-table), make the decision on the basis of p-value w.r.to the given value of beta.

Now a p-value close to zero indicates the variable is significant .

Qn: Explain the Anscombe's quartet in detail:

Ans: Anscombe's Quartet is the modal example to demonstrate the importance of data visualization. This was developed by the statistician Francis Anscombe to signify both the importance of plotting data before analyzing it with statistical properties. To demonstrate this, Anscombe put together 4 data sets that had similar statistical properties (mean, std deviation) when calculated yet when they are plotted, they show different characteristics.

The below is an attempt to run the the analysis on the anscombe sample dataset available in python



```
In [199]: import seaborn as sns
          import pandas as pd
          import matplotlib.pyplot as plt
          anscombe = sns.load_dataset('anscombe')
```

```
In [194]: anscombe.groupby('dataset').describe()
Out[194]:
```

	x								y							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
dataset																
I	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500909	2.031568	4.26	6.315	7.58	8.57	10.84
II	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500909	2.031657	3.10	6.695	8.14	8.95	9.26
III	11.0	9.0	3.316625	4.0	6.5	9.0	11.5	14.0	11.0	7.500000	2.030424	5.39	6.250	7.11	7.98	12.74
IV	11.0	9.0	3.316625	8.0	8.0	8.0	8.0	19.0	11.0	7.500909	2.030579	5.25	6.170	7.04	8.19	12.50

```
In [198]: sns.pairplot(anscombe, hue ='dataset')
          # to show
          plt.show()
```

The code is a few lines as above to illustrate the need for visualization. Essentially, stats show that the mean, std are actually same or similar for all the 4 datasets, but when they are plotted, they are surprisingly different plots.

```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
anscombe = sns.load_dataset('anscombe')
anscombe.groupby('dataset').describe()
sns.pairplot(anscombe, hue ='dataset')
plt.show()
```

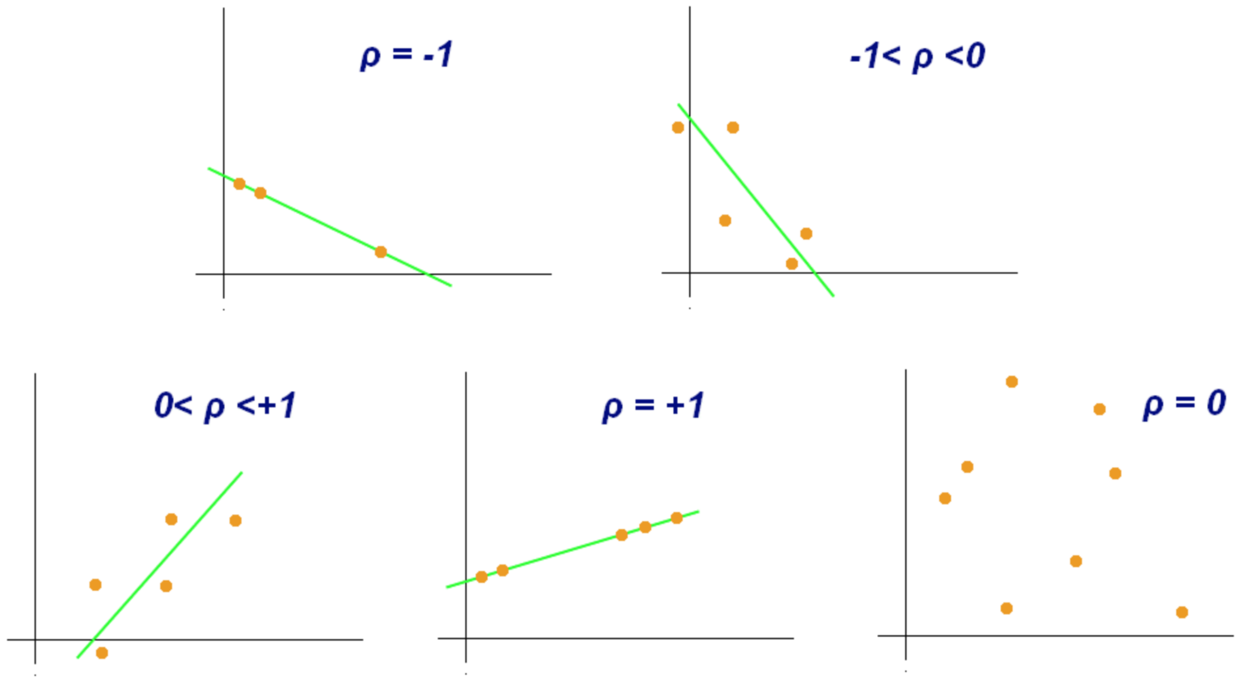
Qn: What is Pearson’s R?
Ans:
It is a statistic that measures the linear correlation between two variables. It has a numerical value that lies between -1.0 and +1.0. Essentially, we can calculate the relationship between two variables by obtaining the value of Pearson's Correlation Coefficient r. Most important criterias for

- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope. A value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient

Below is a quick summary of various possible values for coefficients.



Qn: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans:
Most of the time, the collected data set contains features highly varying in magnitudes, units and range.

If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc

Mainly two Scaling approaches exist

Normalization/Min-Max scaling : In this approach all of the data is brought into range 0 and 1. Sklearn.preprocessing.MinMaxScalar helps to implement normalization.
Standardization scaling: replaces the values by their z scores. It brings all of the data into a standard normal distribution which has mean zero and std deviation one. Sklearn.preprocessing.scale helps to implement standardization in python.

Normalization has a disadvantage of losing the outlier related information once done.

Qn: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans:

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Generally, a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

As the R2 is a measure of the goodness of fit of a model. An R2 of 1 indicates that the regression predictions perfectly fit the data. we get R2 =1, which leads to 1/(1-R2) infinity.

Qn: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans:
Q-Q plots (Quantile-Quantile plots) are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.
If the data is normally distributed, the points in a Q-Q plot will lie on a straight diagonal line. To construct the Q-Q plot, the data values are ordered and cumulative distribution values are calculated as (i– 0.5)/n for the ith ordered value out of n total values. A cumulative distribution graph is produced by plotting the ordered data versus the cumulative distribution values.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively