Assignment-based Subjective Questions

1. ***From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)***

   Answer: Majority of the categorical variables only became driving variables for the model and bike demand. Here is the list of variables which participated in the final model.

   ['**workingday**','windspeed', '**summer**', '**winter**', '**fall**', '**yr**', '**sat**', '**may**', '**oct**', '**mist_cloudy**', '**mar**', '**sep**', '**light_rain**']

   Here except windspeed all other columns are categorical varaiables only. We can infer that based on season and weather conditions are majorly effecting bike demand.

2. ***Why is it important to use drop_first=True during dummy variable creation? (2 mark)***

Answer: When we create dummy variables we use get_dummies method from panda library Like as below.

For ex: season = **pd.get_dummies**(df['season'], drop_first = True)

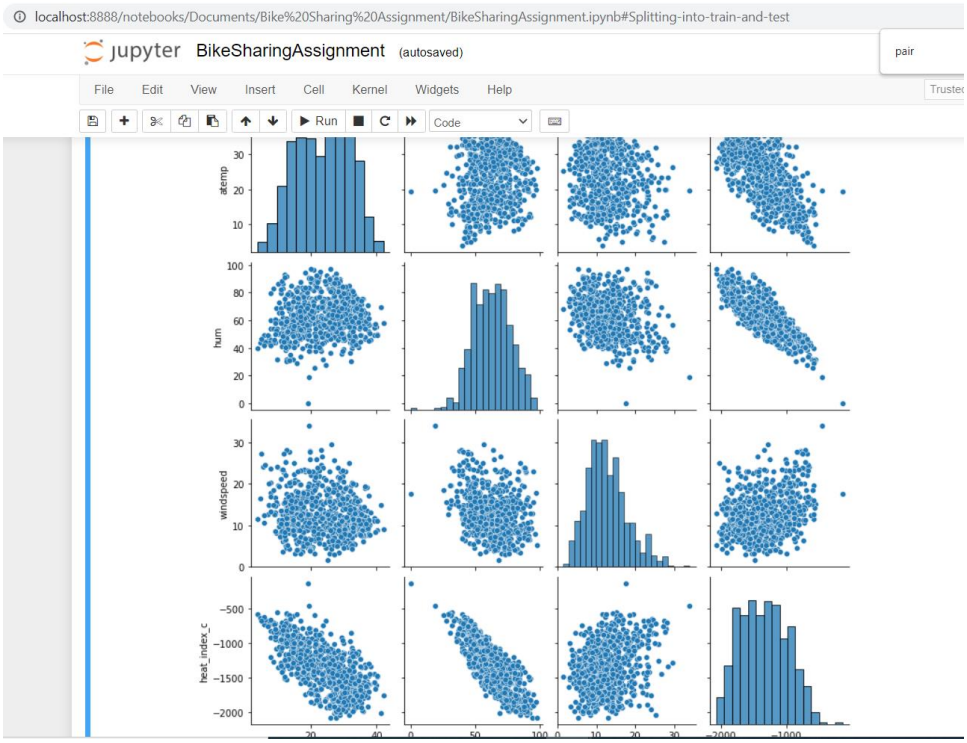
If we have n categories in the column, we will need only **n-1 dummy variables** which other dummy variable can be derived from remaining n-1 dummy variables. This is to avoid more number of features getting created which is going to minimize complexity for the algorithms and even for us to look all the columns would be easy.

3. ***Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?*** (1 mark)

Answer: if we consider **positive correlation, windspeed** has got good correlation. But not to an extent like humidity.

But **humidity is negatively highly correlated**.

This can be observed as shown in picture below.



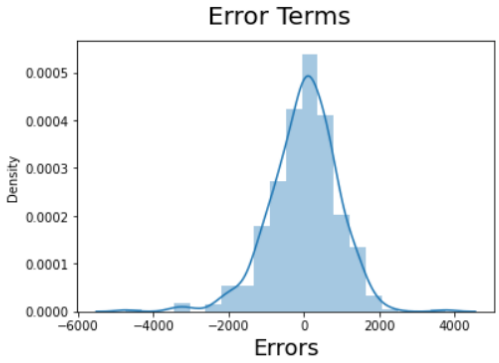
4. ***How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)***

Answer: There are five assumptions we make for building linear regression model.
There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X. An additive relationship suggests that the effect of X on Y is independent of other variables.
There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity. This can be verified with VIF
The error terms must have constant variance. This phenomenon is known as homoscedasticity.
The presence of non-constant variance is referred to heteroscedasticity.
The error terms must be normally distributed. This can be verified with the graph

I have confirmed the assumption of normal distribution of the model as shown below.

```
In [145]: # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((y_train - y_train_cnt), bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)
```

```
Out[145]: Text(0.5, 0, 'Errors')
```



5. ***Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)***

Answer: There is very high negative coefficient (-2771.65) with light rain and its obvious to guess there would not any person who will be interested in bike when it is raining. The next one is positive coefficient with the fall season (2673.16) which is very much convinient for bike riders. So, there is big demand in this season. The next one is positive coefficient with the year (2161.79), which indicates that the second year of the collected information has made more business in terms of bike demand. The next one is summer season (2058.92) has got better business compared other remaining seasons and than other driving variables. Coefficient summary can be seen as below.

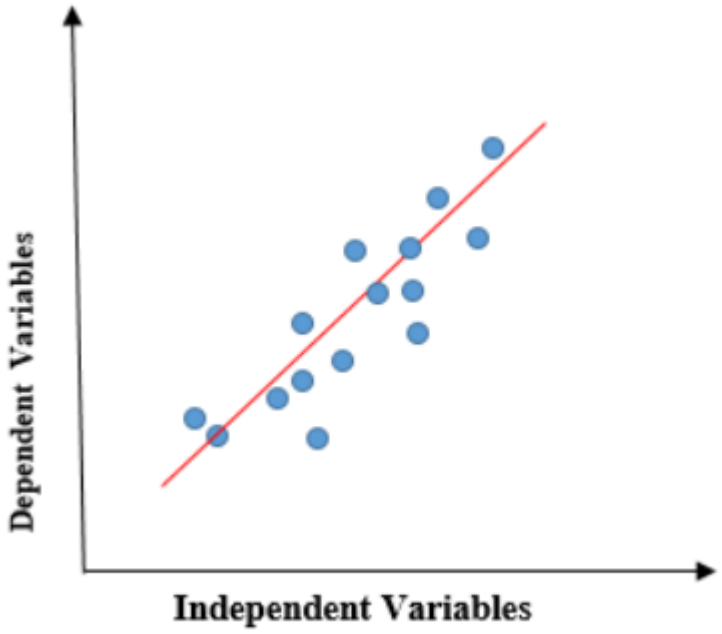
	coef	std err	t	P>|t|	[0.025	0.975]
const	1938.6198	162.794	11.908	0.000	1618.769	2258.471
yr	2161.7909	81.461	26.538	0.000	2001.741	2321.841
workingday	481.8888	111.389	4.326	0.000	263.037	700.741
windspeed	-1467.0530	251.122	-5.842	0.000	-1960.447	-973.659
summer	2058.9222	130.643	15.760	0.000	1802.240	2315.605
fall	2673.1664	129.477	20.646	0.000	2418.775	2927.558
winter	1764.7636	138.206	12.769	0.000	1493.223	2036.305
mar	432.9909	149.430	2.898	0.004	139.397	726.584
may	614.5247	177.039	3.471	0.001	266.687	962.363
sep	717.1337	161.670	4.436	0.000	399.491	1034.777
oct	853.1768	172.824	4.937	0.000	513.619	1192.735
sat	567.4807	143.355	3.959	0.000	285.822	849.139
mist_cloudy	-827.2345	87.067	-9.501	0.000	-998.300	-656.169
light_rain	-2771.6571	247.778	-11.186	0.000	-3258.481	-2284.833

General Subjective Questions

1. ***Explain the linear regression algorithm in detail.*** ***(4 marks)***

Answer: In general, linear regression is very simple statistical regression which is used for predictions and analyze the relationship between continuous or real numbers or numerical variables. It shows the linear relationship between the dependent and independent variables. Here independent variable is represented as X and dependent variable represented as y. Since it is having linear relation between predictor and dependent variables, it is called as linear regression If we have a single variable it is called simple linear regression and if it is multiple or more than one independent variables present, it is called as multiple linear regression. The relation can be positive or negative. Linear regression is supervised category.

It can be mathematically written as **y = mx + c** and can be seen in a diagram as below.



or y = Beta0 + Beta1*x (When there more than one independent variables multiple Bi will become as co-efficients)

In the algorithm, It tries to find best fit line by having minimum Mean Squared Error. It would be used as cost function.

2. ***Explain the Anscombe's quartet in detail.*** ***(3 marks)***

Answer: **Anscombe's Quartet** defined four groups of data sets which are nearly identical in simple descriptive statistics, but there are many peculiarities in the dataset that fools the regression model. They all have very different distributions and appear differently when plotted on scatter plots. These four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

It is used to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. Those four datasets can be described as 1st Dataset which fits the linear regression model properly. 2nd Dataset which could not fit linear regression model on the data quite well as the data is non-linear. 3rd Dataset have the outliers

involved in the dataset which cannot be handled by linear regression model and finally 4th Dataset also have the outliers involved in the dataset which cannot be handled by linear regression model

So, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R? (3 marks)

Answer: Inferential statistics consists of statistical methods that are used to test hypotheses that relate to relationships between variables. For example, you might hypothesize that individuals with greater levels of education tend to have higher incomes. While we can use descriptive statistics such as line plots to illustrate the relationship between these two variables, we need to use inferential statistics to more rigorously demonstrate whether or not there is a relationship between these two variables. With all inferential statistics, which particular statistical test you use will depend on the nature of your data as well as the nature of your hypothesis. When we talk about Pearson's correlation coefficient (also known as Pearson's r), we must understand the chi-square test, the t-test, and the ANOVA. The chi-square statistic is used to show whether or not there is a relationship between two categorical variables. For example, you can use the chi-square statistic to show the relationship between the highest degree completed (e.g., coded as none, high school diploma, bachelors, etc.) and political affiliation (coded as Republican or Democrat). The t-test is used to test whether there is a difference between two groups on a continuous dependent variable. For example, you would select the t-test when testing whether there is a difference in income between males and females. The ANOVA is very similar to the t-test, but it is used to test differences between three or more groups. For example, you would use an ANOVA to test whether there is a difference in income between blacks, whites, and Hispanics. The ANOVA is actually a generalized form of the t-test, and when conducting comparisons on two groups, an ANOVA will give you identical results to a t-test.

Pearson's correlation coefficient (r) is used to demonstrate whether two variables are correlated or related to each other. When using Pearson's correlation coefficient, the two variables in question must be continuous, not categorical. So it can be used, for example, to test the relationship between years of education and income, as these are both continuous variables, but not race and highest degree completed, as these are categorical variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units so, end up in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization(Min-Max Scaling):

- This brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

  MinMax Scaling x = x – min(x) / max(x) - min(x)

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

  Standardisation: x = x – mean(x) / sd(x)

- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: When we have perfect correlation, VIF becomes infinity because we will have R2 as 1 and the calculation for VIF is 1 / 1 – R2. So, because r2 is 1 which makes 1 –R2 as zero. And thus it makes one devided by zero. So, this is example of perfect correlation between two independent variables. For solving this problem, one needs to drop those variables which are perfectly correlated from the dataset. We will call this as multicollinearity. So, this means that corresponding variable can be expressed by a linear combination of other variables. In the given assignment there are multiple highly correlated features available.

They are perfectly correlated to other variables and so, they can be directly removed from participating in the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Merits:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

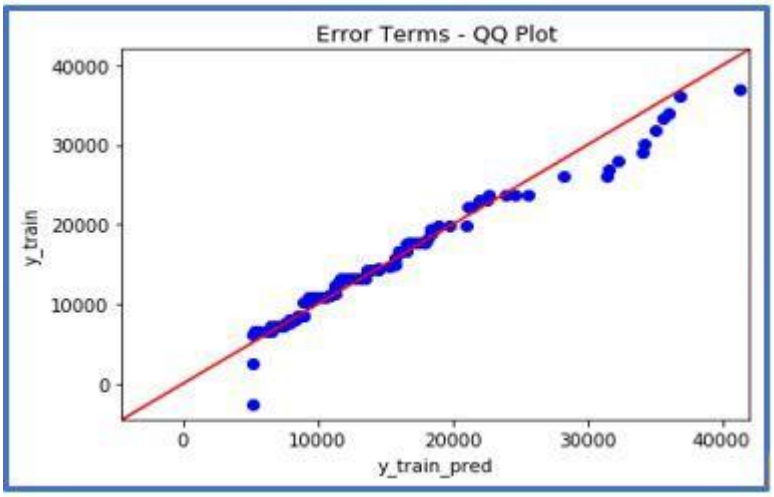
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

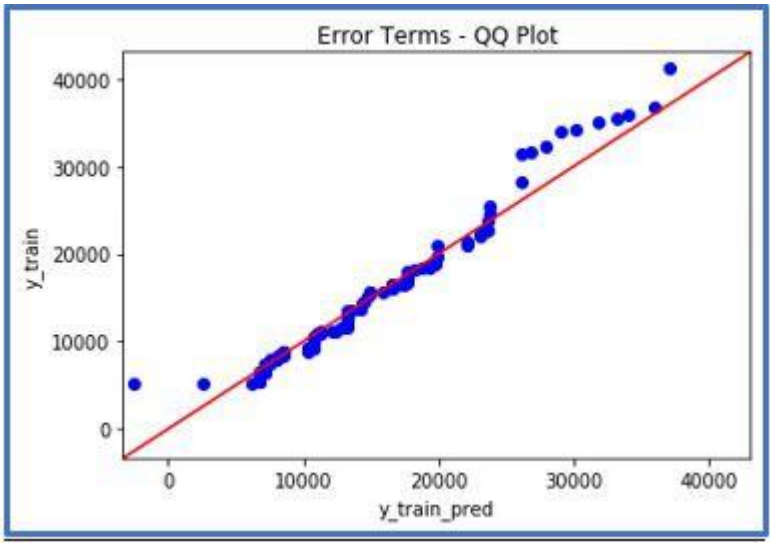
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

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