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)

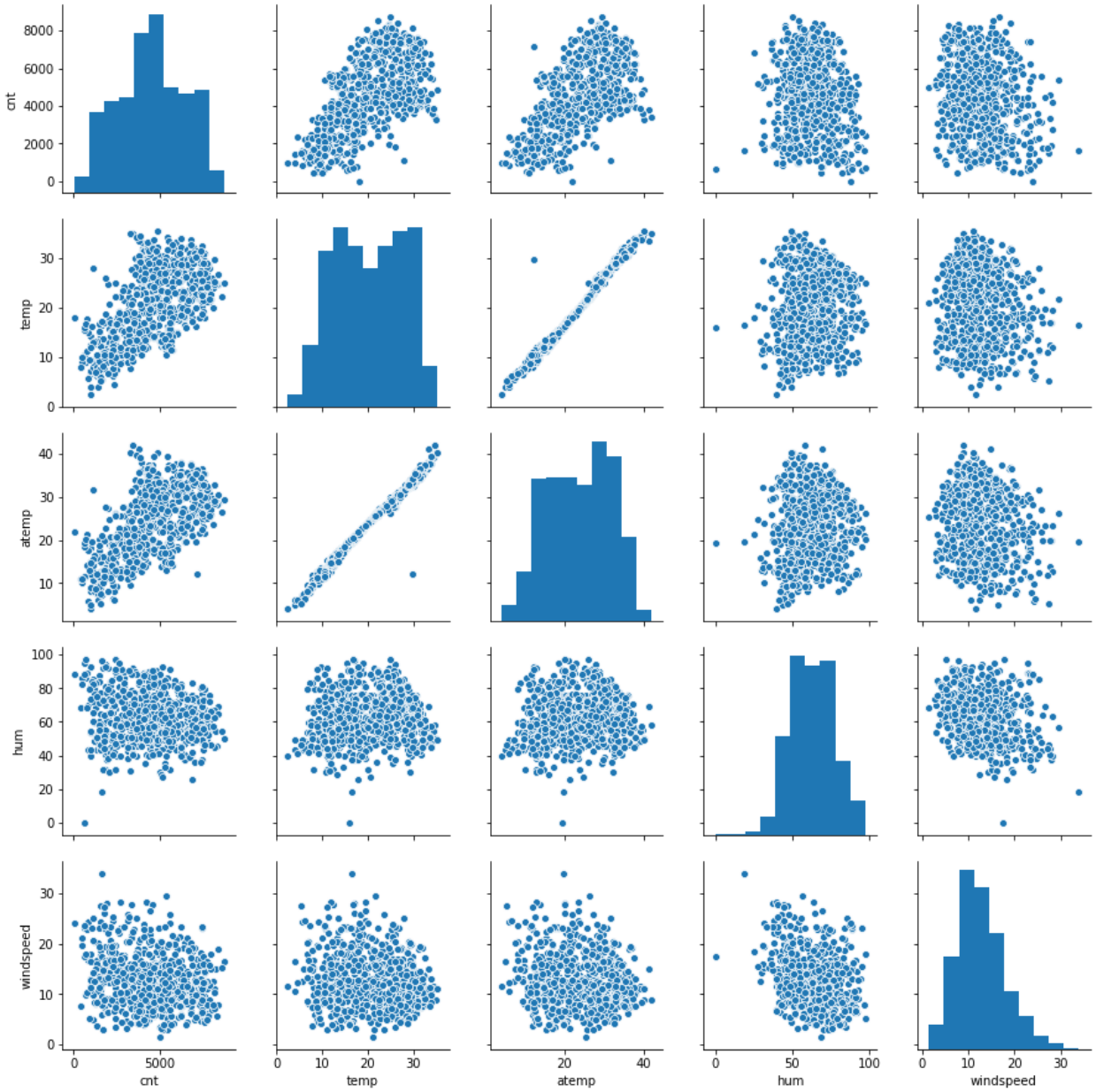
The categorical variable in the dataset were season,yr,mnth,holiday,weathersit and workingday.These were visualized using a boxplot .These variables had the following effect on our dependant variable:-

1. **Season:** The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt.Summer and winter had intermediate value of cnt.
2. **Yr:** The number of rentals in 2019 was more than 2018.
3. **Mnth:** September saw highest no of rentals while December saw least.This observation is on par with the observation made in weathersit.The weather situation in december is usually heavy snow.
4. **Holiday:** Bike Rentals is reduced during holiday.
5. **Weathersit:** There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable.Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
6. **Workingday:** Demand seems to better during non-working days rather than working days for bike rentals**.**

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

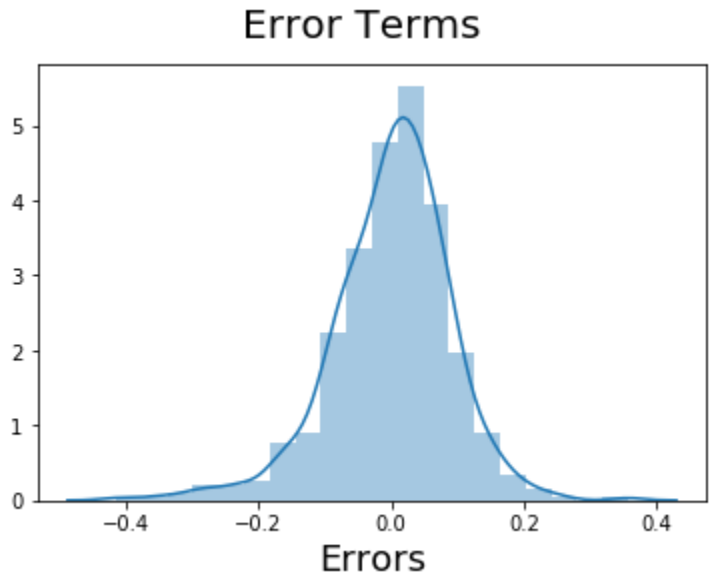
If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importances may be distorted.Another reason is, if we have all dummy variables it leads to multicollinearity between the dummy variables. To keep this under control, we lose one column.

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



"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt)

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



Residuals distribution should follow normal distribution and centred around 0.(mean = 0).
We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.The above diagram shows that the residuals are distributed about mean = 0.

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

The top 3 features are:

1.temp: coefficient value: 0.450735
2. yr: coefficient value: 0.237709
3. weathersit_Clear: coefficient value: 0.096846

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation **"y = mx + c".**

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression : SLR** is used when the dependent variable is predicted using only **one** independent variable.
2. **Multiple Linear Regression :MLR i**s used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots$$

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

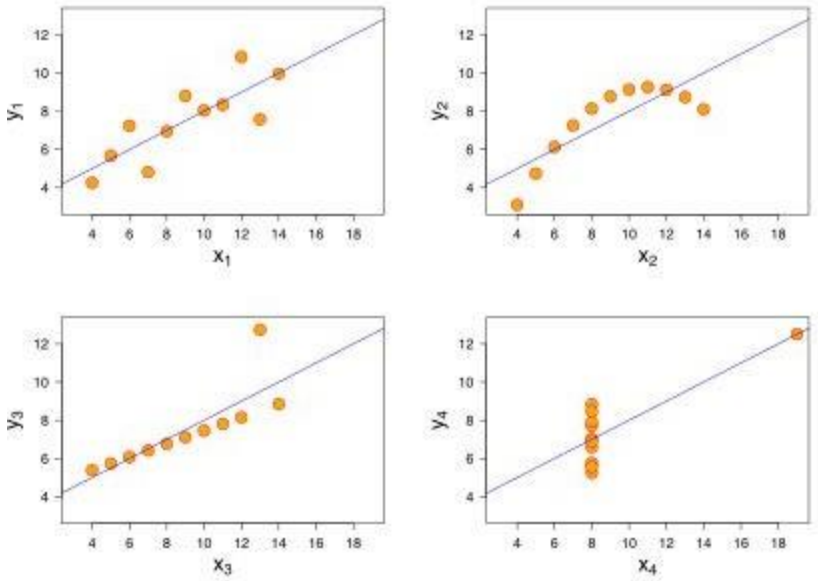
β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.It was developed to emphasize both the importance of

graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data.

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values**.**

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

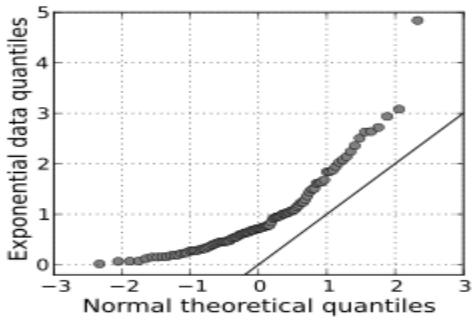
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the QQ plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.