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?**

The categorical variables from the dataset was visualized using boxplot. By looking at the boxplots we are draw some inferences

Season: Spring has the least count has the least bookings. Fall has the maximum, Summer and winter had the intermediate bookings.

Yr: Bookings were higher in 2019 as compared with 2018.

Mnth: Rentals are higher from May to Sepetember.

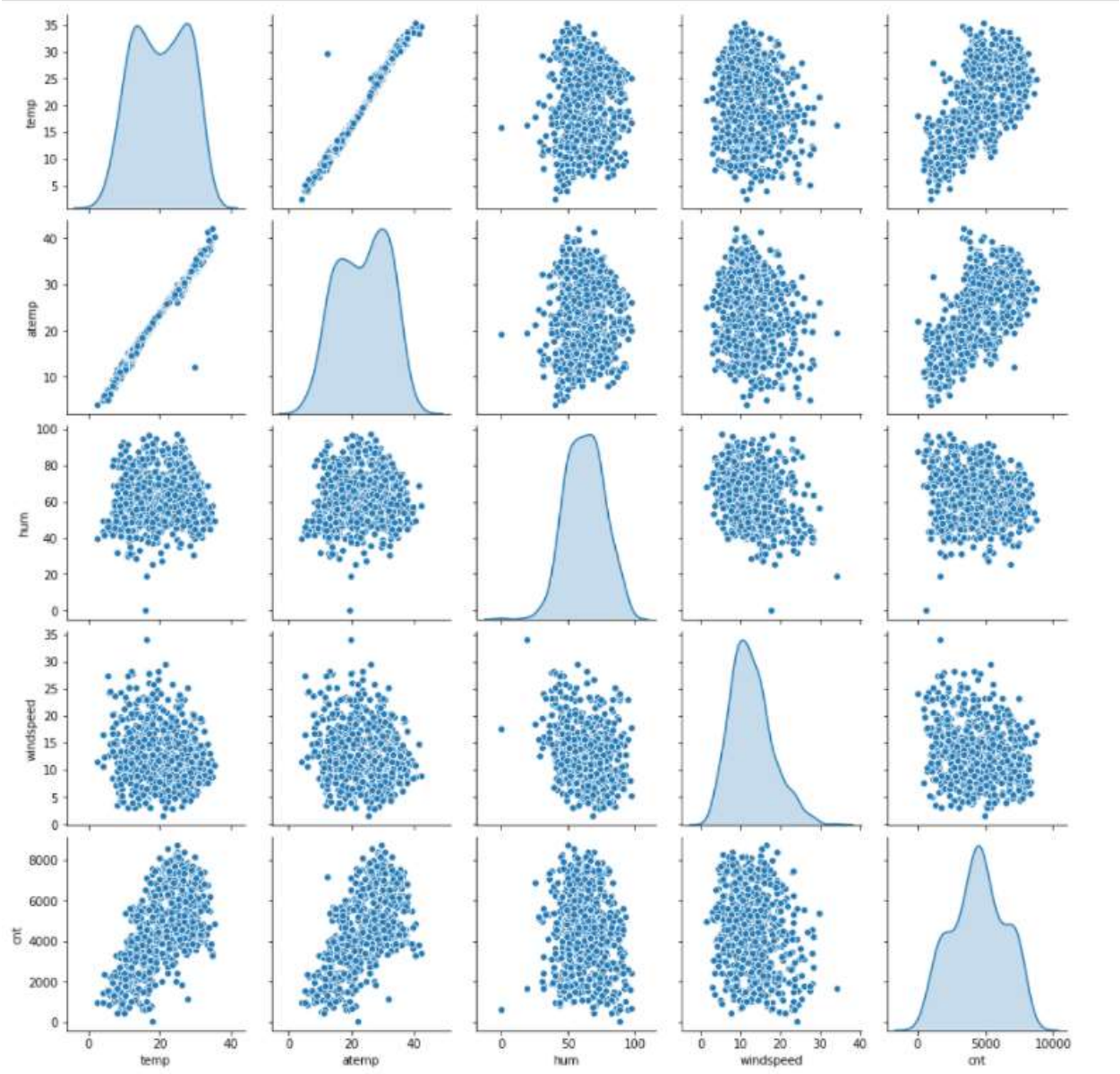
Weathersit: Rentals are higher in Clear+Few clouds. There are no bookings in Heavy Rain+Ice Pallets+Thunderstorm.

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

It helps in reducing the extra column created during dummy variables. It basically reduces the correlation among dummy variables. If drop_first=True is not done, it will have a adverse effect on the models like it may lead to multicollinearity between the dummy variables.

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

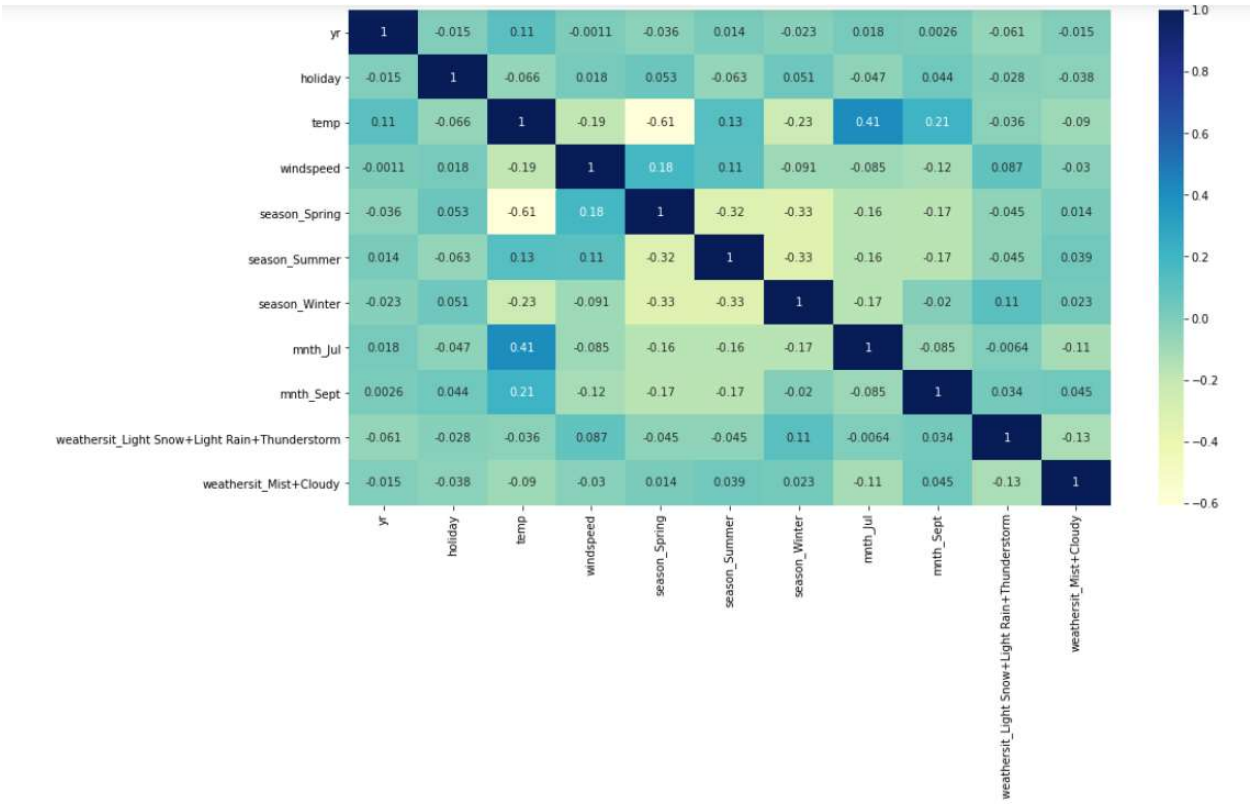
Pair-Plot is created on numerical variables. By looking at the Pair-plot, temp and atemp has the highest correlation with the dependent variable, cnt.



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

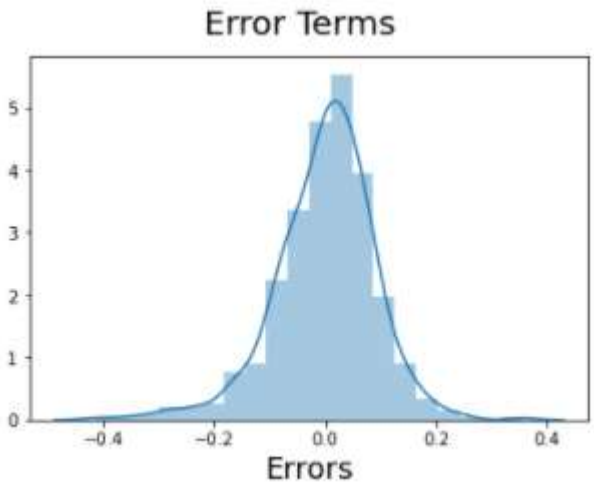
Once the model is built on a training dataset, we consider

1. There should be a linear relationship between dependant and independent variables. This was clear from the heatmap.



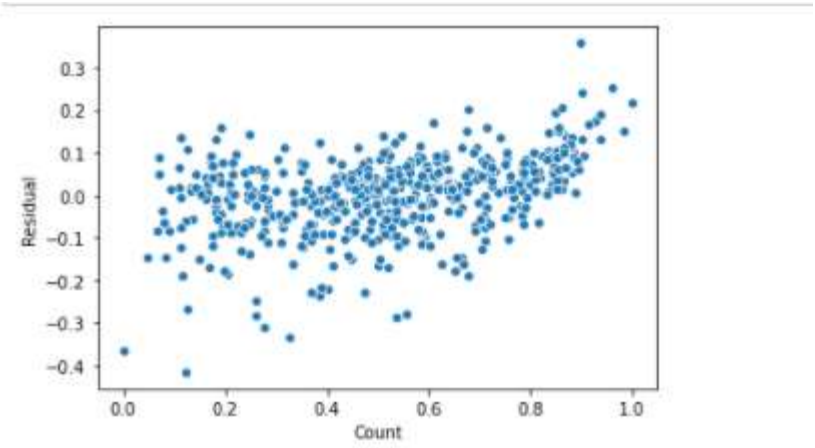
2. Error Terms

Error terms or Residual are normally distributed with mean zero.



3. Homoscedasticity

Error terms have constant variance.



4. Error terms are independent of each other

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

The top three features contributing significantly are

a. Temp coefficient: 0.4915
b. Yr coefficient: 0.2335
c. Weathersit_Light Snow + Light rain+thunderstorm coefficient: -0.2852

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a statistical model used to predict the relationship between dependent and independent variables. Linear Regression is a supervised machine learning technique. Linear regression is the simplest form of regression technique.

In linear regression we find the best-fit line using Ordinary Least Squared method which describes the relationship between dependent and independent variables. It makes prediction for continuous variables like Price, Sales, Age etc. There are two main type of linear regression.

1. Simple Linear Regression

Simple linear regression is used when dependent variable is predicted by using only one independent variable.

Mathematically it is denoted as

$y = B0 + B1x$

where y --> dependent variable
B1 --> slope
B0 --> intercept

2. Multiple Linear Regression

Multiple linear regression is used when the dependent variable is predicted by multiple independent varaibles.

Mathematically it is represented by

$y = B0 + B1x1 + B2x2 + ... + Bnxn$

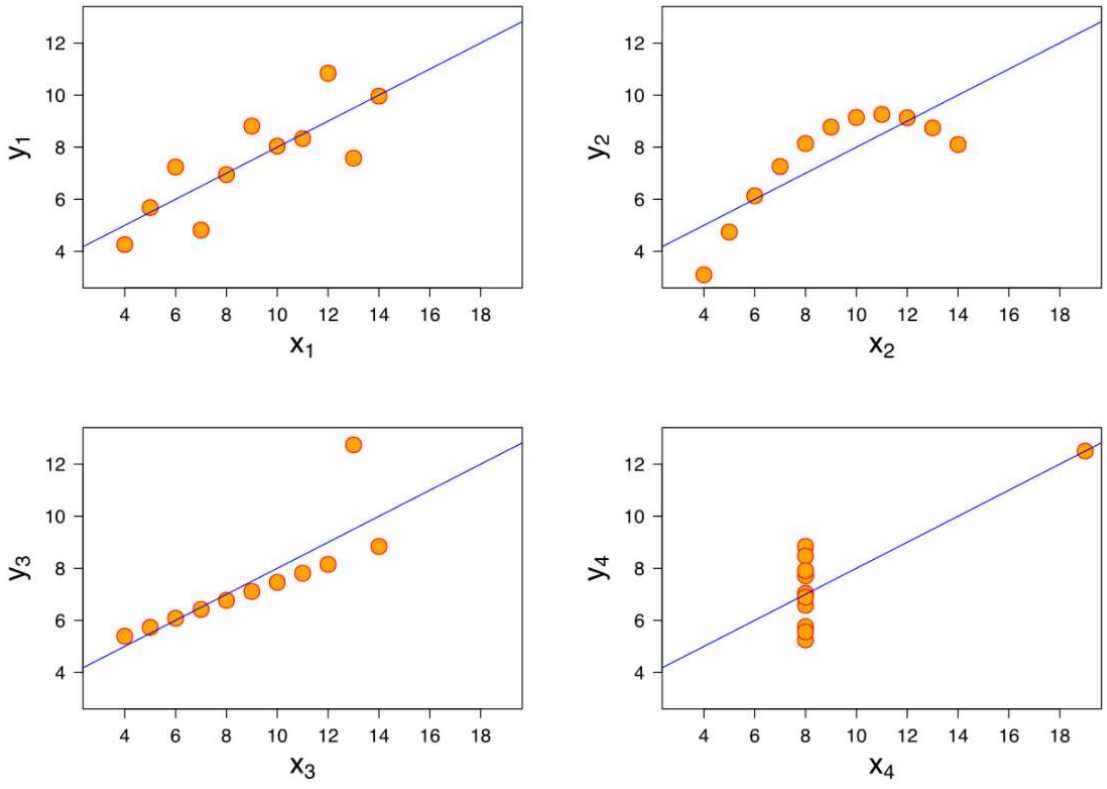
Where B1 --> Coefficient of x1
B2 --> Co-efficient of x2
B0 --> intercept

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

Anscombe's quartet was constructed by a statistician Francis Anscombe in 1973. This was developed to demonstrate the importance of graphical data before analyzing it and the effect of outliers and other influential observations on

statistical properties. It consists of four data sets that have identical descriptive statistics yet have very different distributions and appear different when graphed.

Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.1	10	7.46	8	6.6
8	6.95	8	8.1	8	6.77	8	5.8
13	7.58	13	8.7	13	12.7	8	7.7
9	8.81	9	8.8	9	7.11	8	8.8
11	8.33	11	9.3	11	7.81	8	8.5
14	9.96	14	8.1	14	8.84	8	7
6	7.24	6	6.1	6	6.08	8	5.3
4	4.26	4	3.1	4	5.39	19	13
12	10.8	12	9.1	12	8.15	8	5.6
7	4.82	7	7.3	7	6.42	8	7.9
5	5.68	5	4.7	5	5.73	8	6.9



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

2. The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

3. In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

4. 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)**

The Pearson's correlation coefficient is the test statistics that meausres the statistical relationship between two continuous variables. The correlation between two variables reflects the degree to which the variables are related. It is the most common feature is the Pearson Product Moment Correlaton (also called as Pearson Correlaton) , when measured in population it is denoted by rho and when measured in samples it is denoted by 'r' or Pearson's R. it reflects the degree of linear relationship between two variables. It ranges from +1 to -1.

A correlation of +1 means that there is a perfect positive linear relationship between two variables which means that as the values on X-axis increases, the values on y-axis increases.

A correlation of -1 means that there is a perfect negative linear relationship between two variables which means that as the values on X-axis increases, the values on y-axis decreases.

A correlation of 0 means that there is a no linear relationship between two variables.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Feature scaling is a method used to normalize the range of independent variables or features of data. This is generally performed during pre-processing of data.

Normalization is a technique where we scale a variable to have values between 0 and 1.

Standardization is a technique where we scale a variable based on standard deviation with mean usually 0 and standard deviation as 1.

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

Variance Inflation Factor detects multicollinearity in regression analysis. Multicollinearity is when there is correlation between independent variables in a model.

VIF = 1/(1-R-squared)

If there is a perfect correlation between dependent and independent variables, then R-squared value would be equal to 1 which means there is no residual sum of squares, then VIF becomes infinite.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Quantile-quantile plot is a graphical technique for determining if two samples come from populations with a common distribution. Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Uses: Q-Q plot is used to find if

1. Both the data sets come from populations with a common distribution
2. Both the data sets have common location and scale
3. Both the data sets have similar distributional shape
4. Both the data sets have similar behavior

Importance:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.