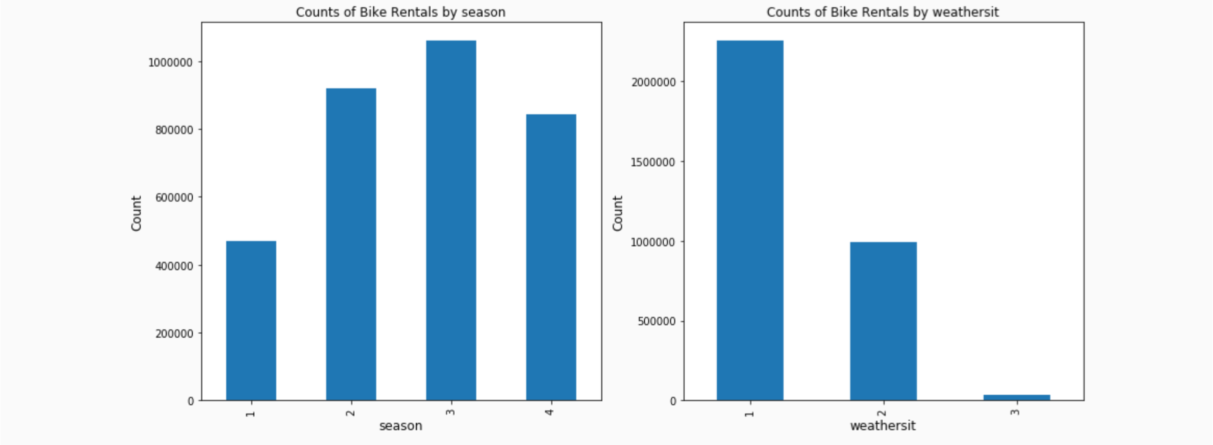
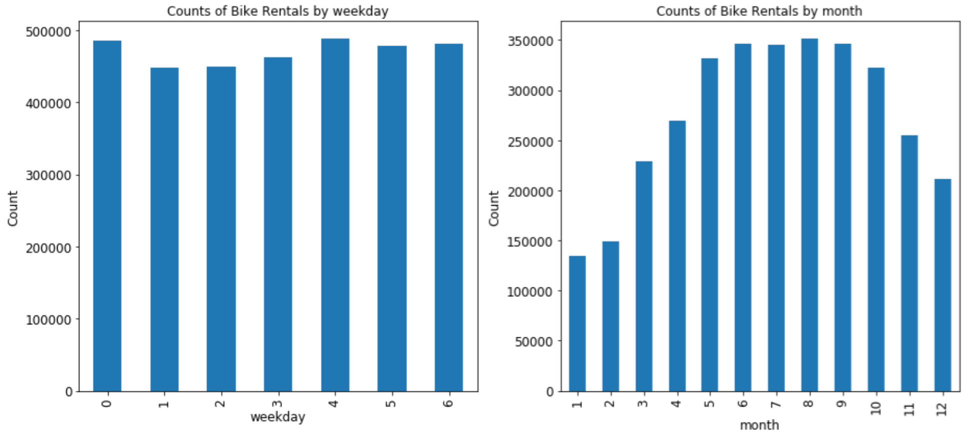
A) Assignment-based Subjective Questions

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





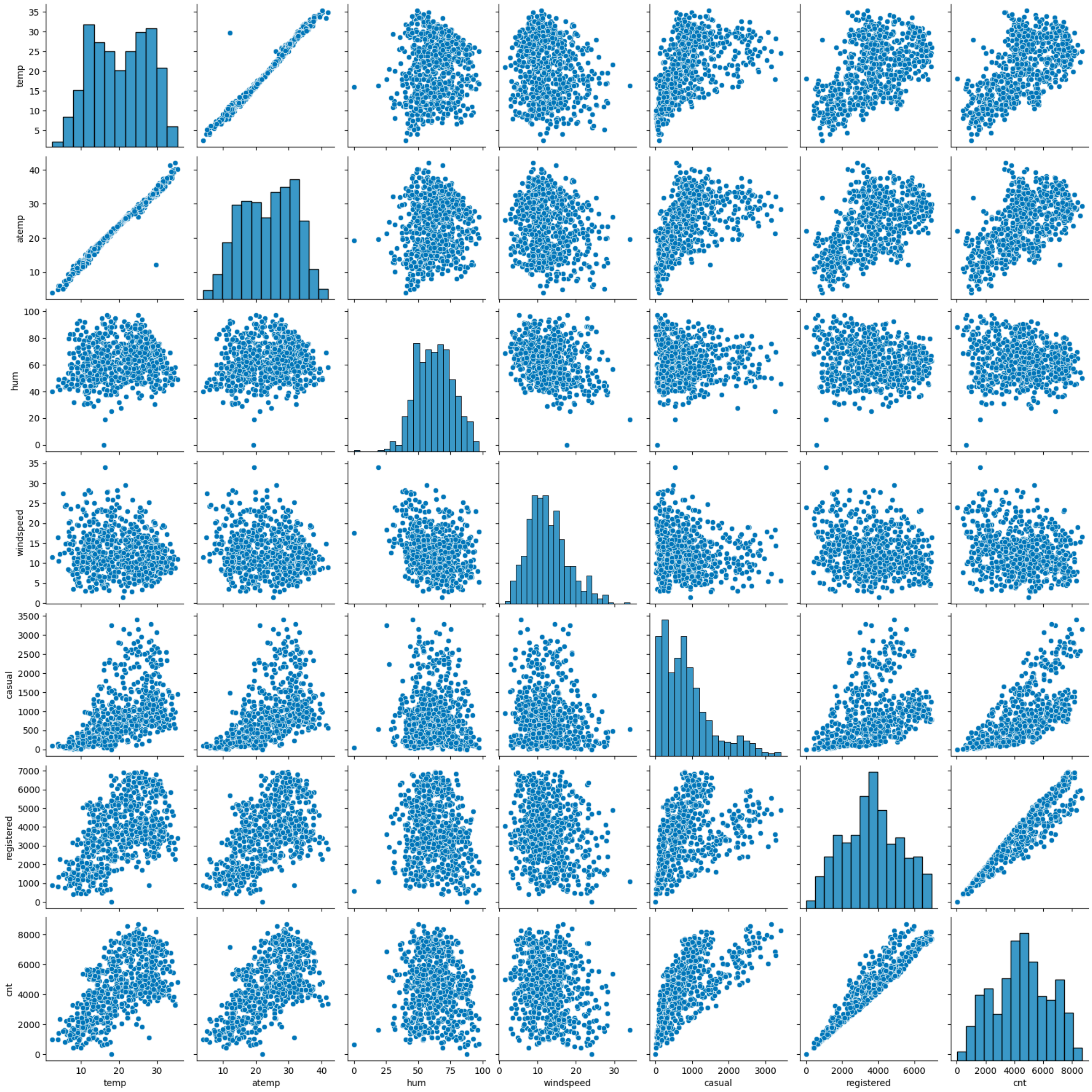
Number of bikes rented is high:

1.Season- Fall is the Top season where the number of bikes rented is high.
2-Weather- The number of bikes rented is high when the weather is clear , few clouds
3.Weekdays-The number of bikes rented goes high during mid week.
4. Month--The number of bikes rented goes high during mid year.

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

If you don't drop the first column then your dummy variables will be. This may affect some models adversely and the effect is stronger when the cardinality is smaller. Hence, Drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished

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



Observation :

1. Atemp and temp have a linear relationship with the target variable.
2. Windspeed and humidity don't appear to have any relationship with the count variable.

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

We can validate the assumptions of Linear Regression after building the model on the following training set by below method:
1)Fitted regression line is linear.
2)Error terms came out normally distributed with mean as 0.

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

The changes of increasing the number of bikes being rented increases during the working day. The demand for bikes on rent if negatively affected by windspeed. The demand for bikes on rent is high in Fall season. The demand of bikes on rent is high in clear weather.

Significant variables to predict the demand for shared bikes

- holiday
- temp
- windspeed
- Season
- months
- Year (2019)
- weathersit

B) General Subjective Questions

1. Explain the linear regression algorithm in detail.?

Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.

Linear regression is of the 2 types:

i.) Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:
$Y=\beta 0+\beta 1X1 +\epsilon$

ii.) **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression:
$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$

The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by the following two methods:

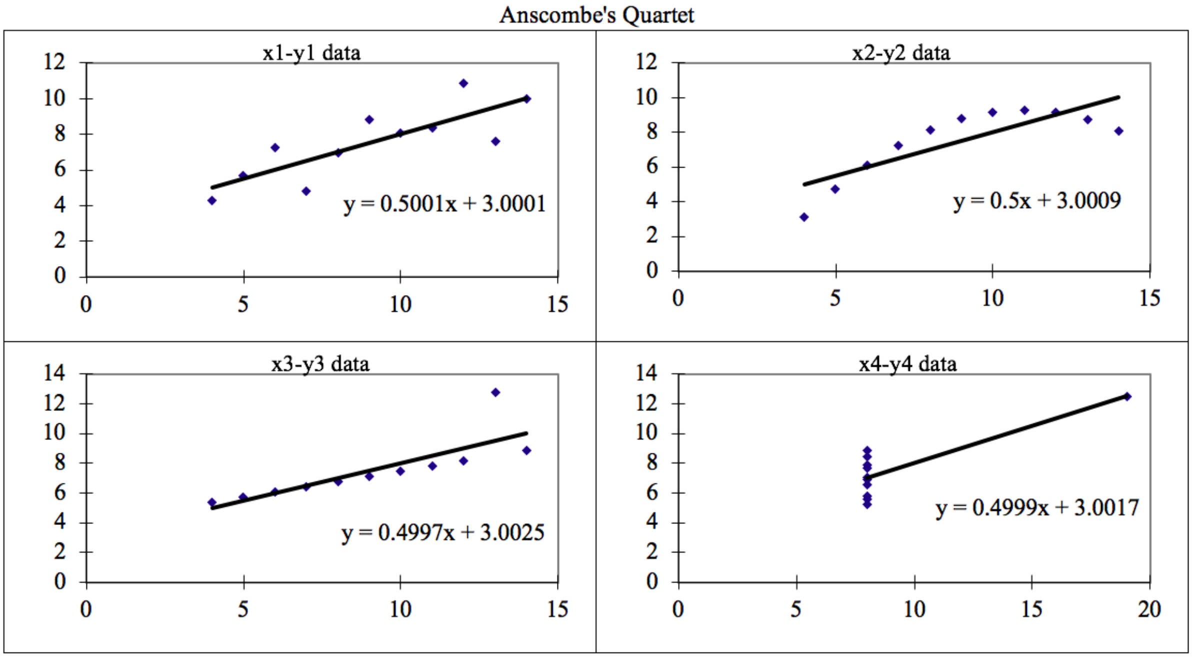
i.) Differentiation
ii.) Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression.

2. Explain the Anscombe's quartet in detail?

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

For example:



3. What is Pearson's R?

The Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed

- The association should be linear
- There should be no outliers in the data

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

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 hence 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: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling: It 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). sklearn.preprocessing.scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it lose 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?

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly 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.

Few advantages:

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.