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?

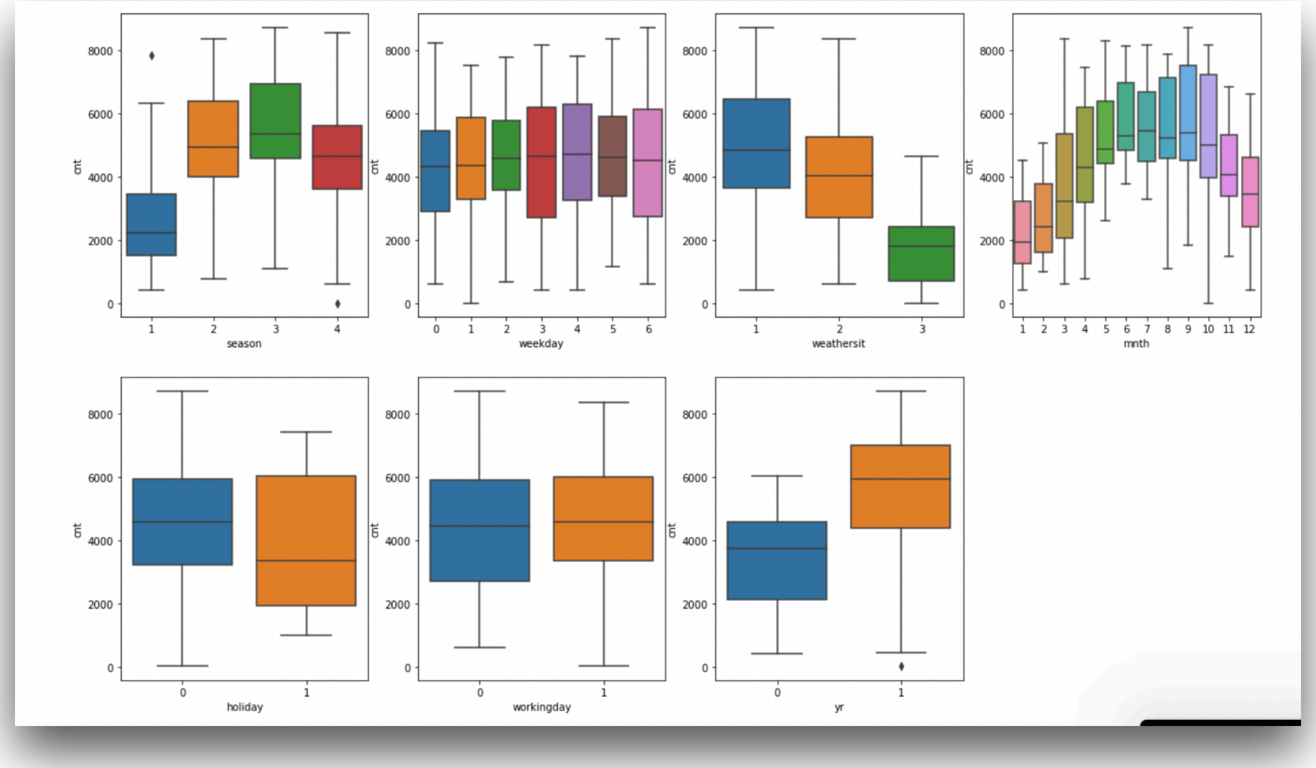
Answer:

The categorical variables that are in this analysis are season, weekday, month, weatherise, holiday, and working day

Used box plot to visualise these categorical variables in the analysis. The inferences that can be derived on bike bookings from these are:

- Season: Highest number of bookings can be seen in Fall and next highest booking season is Summer. Whereas, lowest number of bookings can be seen in Spring season. According to the bivariate analysis number of books increased in all seasons from 2018 to 2019
- Weekday: This doesn’t see a significant trend between each weekday. However, could see increase in number of bookings. Medium remains almost same in all the weekdays
- Weathersit: Bookings are higher in case of “Clear, Few clouds, Partly cloudy, Partly cloudy”. However, could see lowest number of bookings in “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds”. There are no bookings at all in “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog”.
- Month: September see largest number of bookings. This is consistent with the model inference at the end. Lowest bike bookings could be seen in December and January. As seen about least bookings are due to snow. This season dip also explains that!
- Holiday: If it is a holiday the number of bookings reduce. This is consistent with final model inferences.
- Workingday: This doesn’t have an impact on bike bookings
- Year: As compared to 2018, number of bookings increased for all the above mentioned scenarios in 2019. Overall number of bookings increased by more than 2000 in 2019 as compared to 2018.

Below are the box plots:



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

Answer:
While creating dummy variables for “n” categorical variables, we need n-1 dummy variables. For example if there are two categorical variables “male” and “female” for gender column. Then we need 1 dummy variable to identify the gender, that is ‘0’ for male and ‘1’ for female.

While creating dummy variables using pandas in python we use drop_first=True to create n-1 dummy variables and remove the redundant variable or highly correlated variable. Without using this “n” dummy variables are created for n categorical variables and leads to multicollinearity and the derived inferences will be unreliable.

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

Answer:
The numerical variables “temp” and “atemp’ have highest correlation with target variable “cnt”

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

Answer:
To validate the assumptions of Linear Regression after building the model on the training set are:

- Linear relationship between variable and target variable
  - This is achieved by plotting scatter plot between test and prediction of target variable
- Multicollinearity between features should not be there
  - There should be no high collinearity between variables
- Error terms should be normally distributed
  - This cam be verified by plotting distplot between error terms and density
- Homoscedasticity
  - Error terms should not be same across all values of independent variables.

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

Answer:
Top 3 features contributing significantly towards explaining the demand of the shared bike are:

- Temp - 0.568
- Year - 0.233
- Windspeed - -0.145

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:
Regression is a supervised machine learning algorithm. That means machine learns or trains from the previous data with labels that can be used to develop predictive analysis model.
Various industries such as economics, finance, engineering, medicine, sports and entertainment use these predictive analysis models to predict the future outcomes of their business

Linear regression is one of the predictive analysis model, which is used to know the relationship between target variable that is dependent variable and predictors that are independent variables.

There are two types of linear regressions:

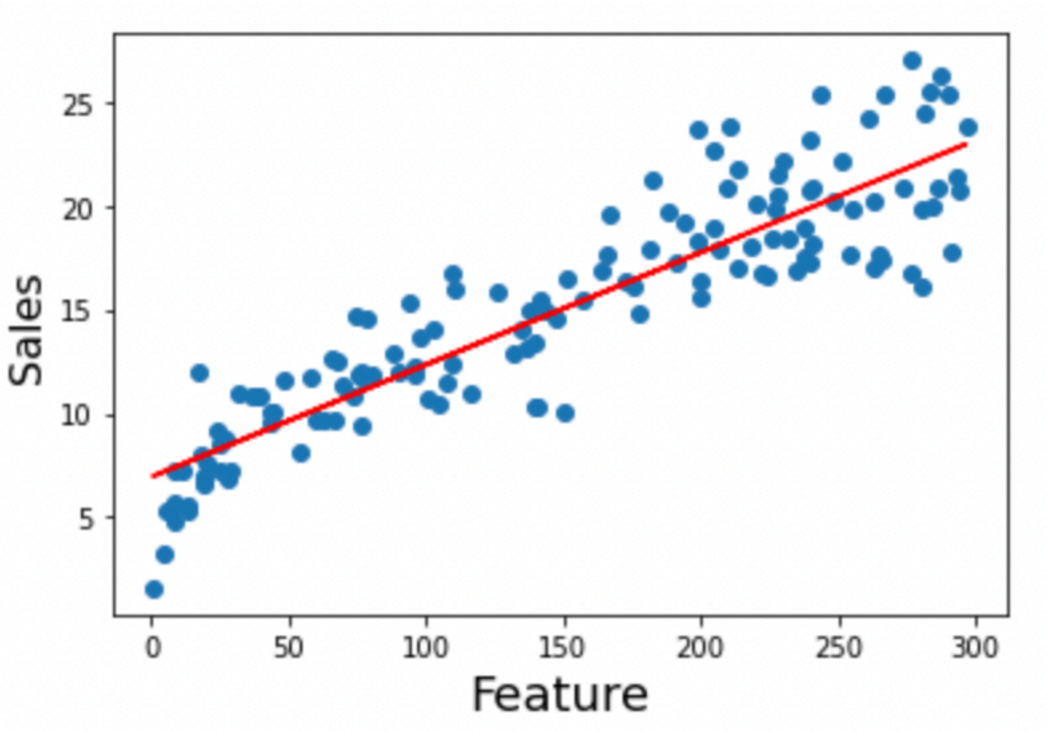
- Simple Linear Regression (SLR)
- Multiple Linear Regression (MLR)

Simple Linear Regression:

As the name suggests this is the basic regression model which provides relationship between a dependent variable to that of one independent variable. The relation should be a straight line. This can be checked by plotting scatter plot between these two variables.

The mathematical equation of simple linear regression is: $y = c + mx$

Where, c is intercept
m is slope, that is, coefficient of feature x
y is the dependent variable



The strength of a linear regression model is defined by below two ways:

- R-Square - This explains what percentage of variance is explained by the developed model. This lies between 0 to 1. Higher R-square value specifies how good your model fits the data.
- Residual Standard Error (RSE) - This is the difference between expected and actual output. Lesser the RSE indicates better fit the model.

Multiple Linear Regression

This is a regression model which provides relationship between one dependent variable and multiple independent variables. The linear equation the best fit the dependent variable Y with multiple independent variables X

The mathematical equation is: $\mathbf{y = c + m_1x_1 + m_2x_2 + \ldots\ldots + m_nx_n}$

Where y is target variable, c is intercept and m1 is coefficient of feature x1 ans so on

New considerations that are required while shifting to multiple linear regression, are:

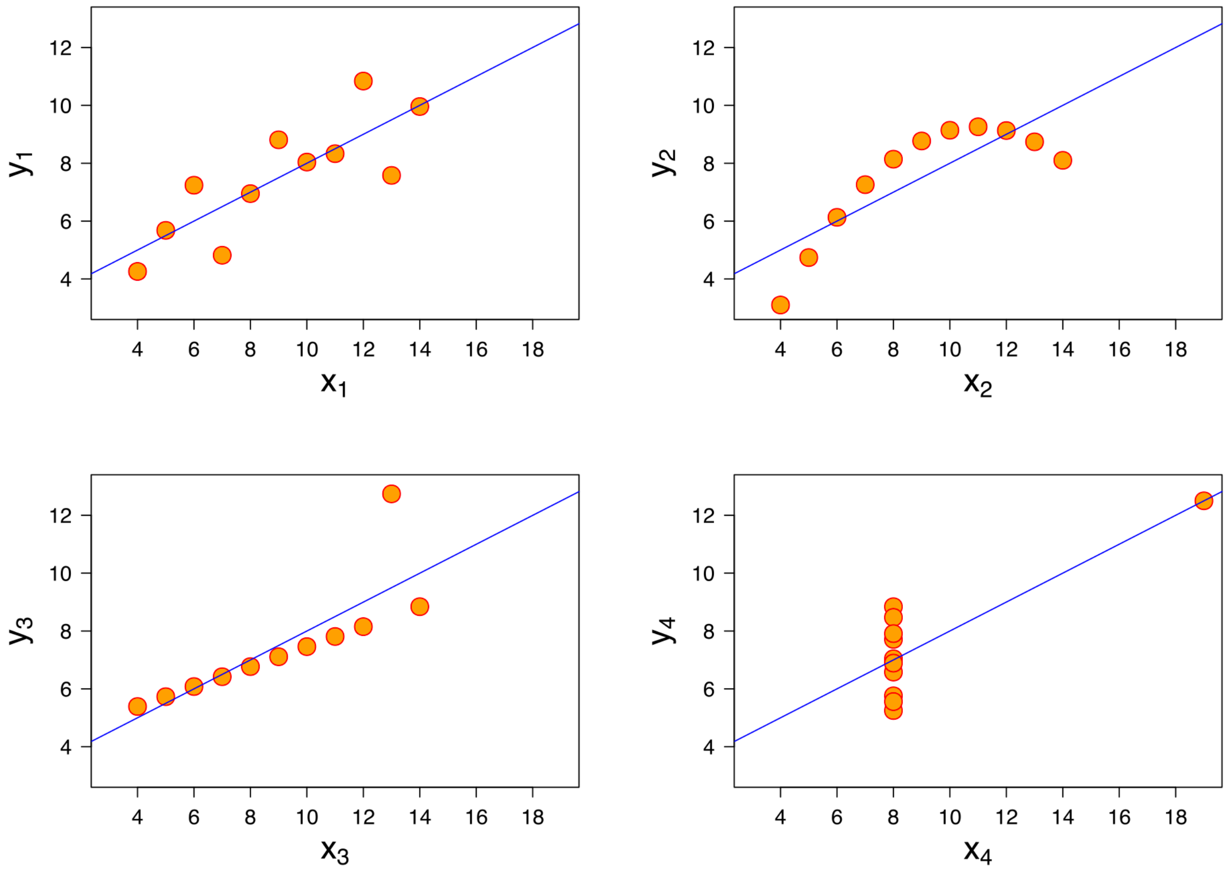
- Overfitting - Adding more features may have negative side-effect
  Model fits the train set 'too well', it memorises are data-points and doesn't perform well on test-set
- Multicollinearity - High association between predictor variables

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet as name suggests consists of four data sets with eleven points (x,y) that have almost similar descriptive statistics, however, they have very different statistics and appear different when plotted in a graph.
This quartet is constructed in 1973 by statistician "Francis Anscombe", to mainly demonstrate the importance of plot graph while analysing it and also shows the effect of outliners and other influential observations on statistical properties. The major ideas behind this is to counter the impression that "numerical or theoretical calculations are accurate and graphical representation is unnecessary" which generally statisticians of that time have.

The four data sets are identical when analysed numerically, however, different when plotted on a graph, as shown below:



[Graph source from: https://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg]

The x1 vs y1 plot seems to be a simple linear regression
The x2 vs y2 plot , the data points seems to be non-linear
The x3 vs y3 plot the data points are linear, however, the regression line is different from this data points because of the influence of one outliner
The x4 vs y4 plot the data points do not show any relationship between variables, however, the regression line is different due to high leverage point.

Hence it is recommend to graphically plot data before starting to analyse.

3. What is Pearson's R?

Answer:
Correlation is measure of relationship between sets of data The most common measure of correlation in stats is Pearson Correlation which is developed by Karl Pearson. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data.
There are two letters to represent the Pearson correlation: Greek letter rho (ϱ) for population and the letter 'r' for sample.

This is the ratio between covariance of two variables and product of their standard deviation. The result is always between -1 and 1

- 1 means a strong positive relationship - If there is positive increase in one variable then the other variable increases by a fixed amount positively
- -1 means a strong negative relationship. - If one variable increases positively there is a fixed amount of decrease in other variable
- 0 means no relationship at all. - The two variables are't related

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

Answer:
Scaling is an important feature when there are lot of independent variables in the model. Because, all the independent variables would be generally be in different scales and without scaling the model will not give comparable results. The result will be influenced not by considering the unit of variables but based on the value of variable. Which is not the correct interpretation. Hence, it is advisable to use either normalisation or standardisation to obtain the variables to the same scale.

For example: 1000 meters is given higher value what 2 km, which is undesirable. Hence scaling is used to get these values to a known range. In that way it helps in accurate interpretation of data set

Scaling is performed for ease of interpretation and faster convergence of gradient decent methods

Difference between normalised scaling ans standardised scaling are:

Normalized	Standardised
This scales values to range [0,1]. We call it MinMax scaling	This scales values to have mean 0 and standard deviation 1. We call it Standardizing

Normalized	Standardised
$x = \frac{x - mean(x)}{sd(x)}$	$x = \frac{x - min(x)}{max(x) - min(x)}$
This is effected by outliners	This is less effected by outliners
Scikit library uses MinMaxScaler	Scikit library use StandardScaler

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

Answer:
VIF means Variance Inflation factor. VIF helps in providing information of relationship of one independent variable with all other independent variables.
VIF = 1/1-R^2
When R = 1, then VIF will become infinity
Interpretation of VIF value:

- VIF > 10 is definitely high and variable should be removed.
- VIF > 5 should also not be ignored and inspected appropriately.

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

Answer:
The Q-Q plot means quantile-quantile plot, which is a plot of quantiles of one data-set to other to see if there are any similarities in shapes of distribution.
If the data sets have similar kind distribution then the line turns out to be straight at an angle of 45 degrees.

Importance of Q-Q plot in linear regression is in scenarios where training and test data set are received separately and by using Q-Q plot we can confirm that both the data sets are from populations with same distributions.

Uses of Q-Q plot is:

- To detect the presence of outliners
- It is used to identify if two data sets come from populations with common distribution, have same location and scale, have same distributional shapes and same behaviour.