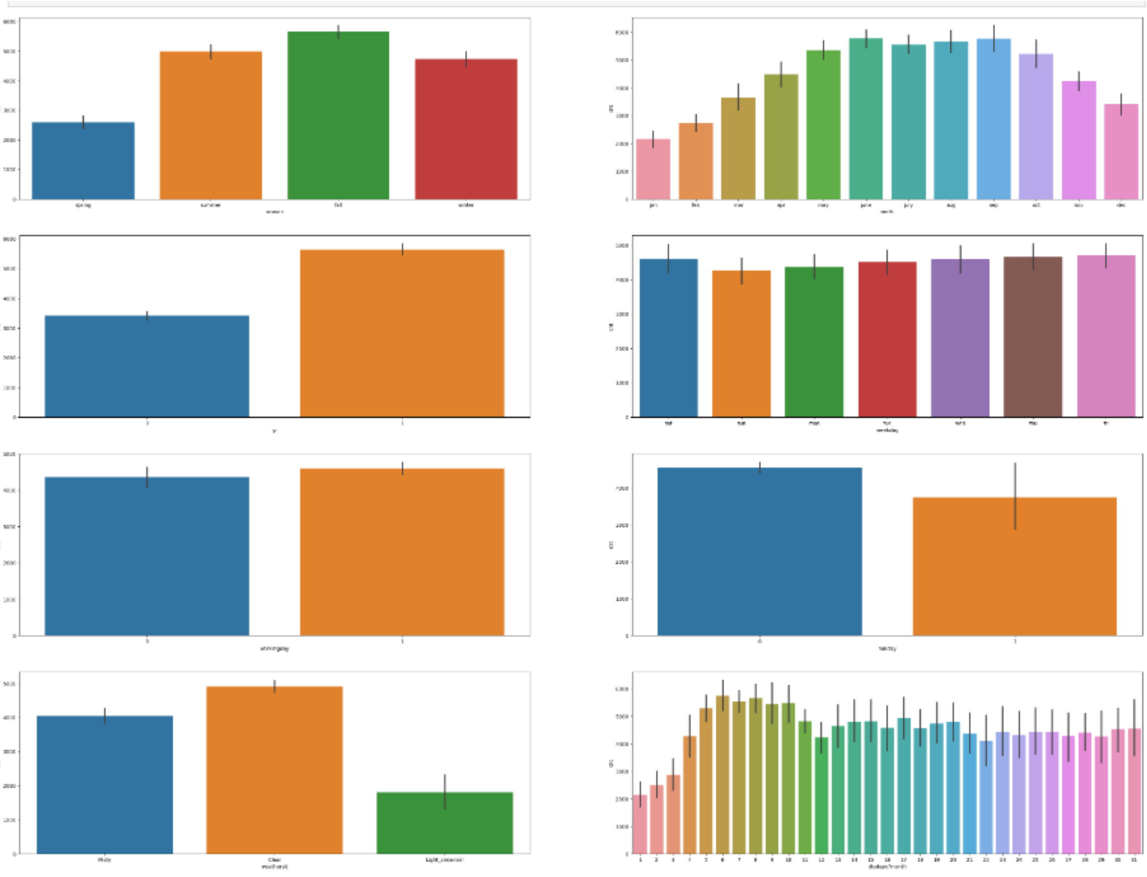
Assignment-based Subjective Questions

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

Ans: Based on the analysis for categorical variables, below are the observations. The same can be seen in bar charts below

- Fall season seems to have attracted more booking. Also with the change in year, for every season, count of bookings got increased
- 2019 attracted more number of bookings as compared to 2018, which shows growth in business.
- In a year, the bookings got gradually increased till June and started to decline till December
- Clear weather attracted more bookings followed by misty weather
- Booking seemed to be almost equal either on working day or non-working day.
- Days staring from 4th to 11th seem to have higher number of bookings as compared to other days of the month



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

Ans: If we do not use drop_first = True, then n dummy variables will be created, and these variables are correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

Ans: Temp variable has highest correlation with the cnt variable

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

Ans: Based on below mentioned parameters the validation was made

- Multivariate normality: There should be insignificant multicollinearity among variables.
- There should be no visible pattern in residual values.
- No auto-correlation
- Linear relationship : Linearity should be visible among variables
- Homoscedasticity : The variences in y_test and y_test_pred was similar

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

Ans: Three features are

- Workingday
- Season(spring)
- Months(Jan)

General Subjective Questions

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

Ans: Linear regression is a form of predictive modelling technique which analyses the relation between dependent variable and set of independent variables defining it. The change in independent variable can impact the dependent variable positively or negatively.

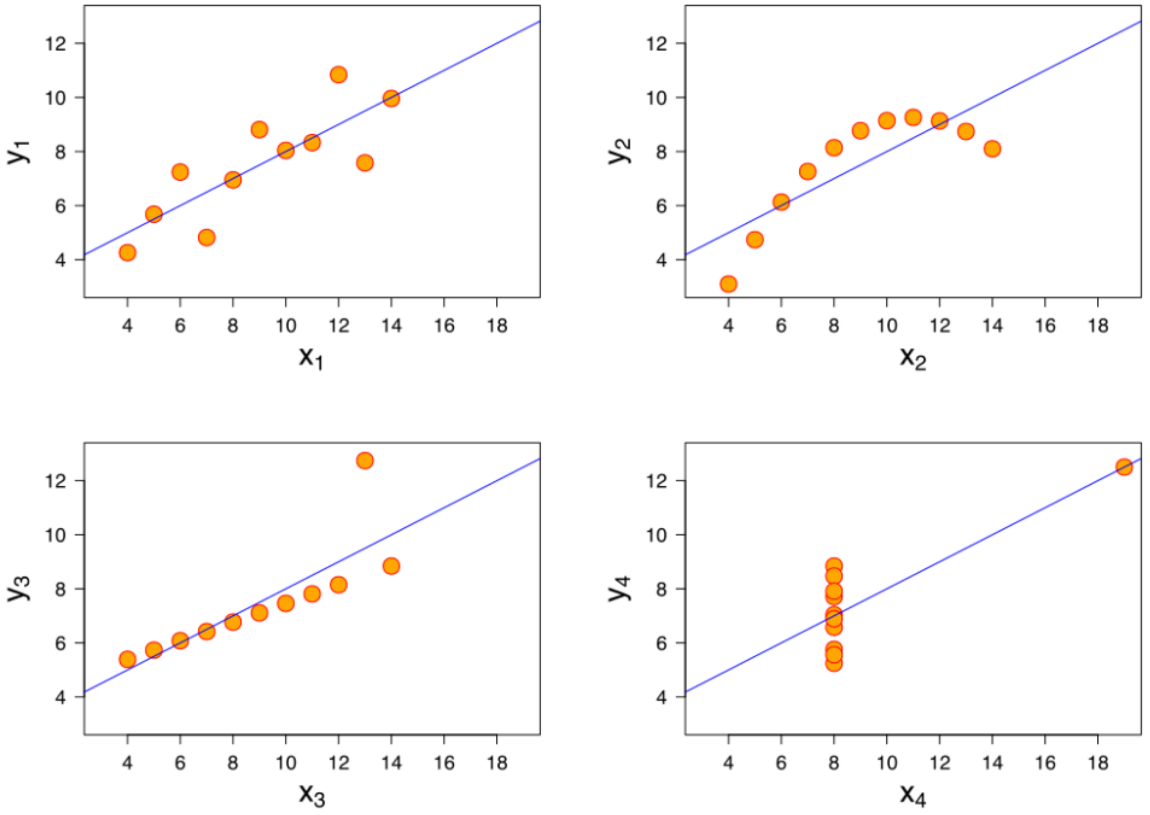
It can be represented by below equation

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots$$

Where x's are independent variables and y is a dependent variable

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

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.



- 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.
- The second graph (top right); 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.
- In the third graph (bottom left), the modelled relationship 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.
- 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)

Ans: This is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- Positive correlation (Between 0 and 1) : When one variable changes, the other variable changes in the **same direction**.
- Zero : There is **no relationship** between the variables.
- Negative correlation (Between 0 and -1) : When one variable changes, the other variable changes in the **opposite direction**.

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

Ans: Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.

2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.

3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.

4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.

5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal. 6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization

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

Ans: VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables A large value of VIF indicates that there is a correlation between the variables.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the used dataset.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The advantages of the q-q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale

- If both datasets have similar type of distribution shape
- If both datasets have tail behavior