Submitted by: Merrymel George

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?

   Listing down the different categorical variables in the dataset along with the inference:
   a. Season – Fall has the maximum booking followed by Summer, Winter while Spring has the minimum. So, it may be inferred that Spring might have a negative effect on the dependent variable
   b. Year – There is an increase in the number of bookings from 2018 to 2019. This should mean that year should be positively affecting the bookings count.
   c. Month – The average count of bookings shows an increase from January till June and then starts to dip down. This could mean that there might be certain months that should be influencing the bookings
   d. Holiday – The demand is less on holidays, and hence it should be having a negative effect on the target variable
   e. Weekday – The median of bookings count remains almost at the same value for all the weekdays. Therefore, it's unclear how weekdays will impact the dependent variable
   f. Working day – Working day variable also has a constant median for both working and non-working days. And so, similar to the weekday variable, it's difficult to infer the effect
   g. Weather situation – Generally clear weather attracts more demand.

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

   During dummy variable creation using encoding, as the first column becomes the reference variable for encoding the rest columns, there are high chances that the dummy variables become correlated. Also, for an n level categorical variable, n-1 columns are enough to

represent the data as one of the n columns can be deduced from other n-1 columns. As such, drop_first=True helps in dropping the unwanted column while dummy variable creation and thereby solves the mentioned issues.

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

   After initial analysis and data cleaning, the numerical variables that are remaining as temperature, humidity and windspeed. Among these 3, temperature has the highest correlation with a positive effect on the target variable

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

   a. Residual Analysis
   b. Confirmed that the residual data is normally distributed and has a mean equal to 0
   c. Confirmed that there are no patterns between residues and predicted values
   d. Confirmed the homoscedasticity – constant variance of the error terms
   e. Confirmed the linear relationship between the independent variables to the dependent variable
   f. Made sure there is no multicollinearity and overfitting by checking the VIFs and R Squared-Adjusted R Squared values
   g. Also, checked for Mean Squared Error value, Root Mean Square Error and Mean Absolute Error value are in the permissible ranges

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

   a. Temperature with the highest positive coefficient
   b. Weather Situation = 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) has the next highest coefficient in magnitude. But the sign is negative

c. Year has the third-highest effect, and it is positive

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Regression analysis is a technique of predictive modelling that helps to find out the relationship between scalar response and one or more explanatory variables (also known as dependent and independent variables). The output variable to be predicted is continuous. Regression guarantees interpolation of data and not extrapolation.

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a
   regression task. Regression models a target prediction value based on independent variables. It is mostly
   used for finding out the relationship between variables and forecasting. Linear Regression is a form of parametric regression

   There are two types of regression–
   1. Simple linear regression: Model with one independent variable. The model attempts to explain the relationship between a dependent and an independent variable using a straight line.
   2. Multiple linear regression: Model with more than one independent variable.

   Assumptions of Linear Regression –
   1. There is a linear relationship between X and Y
   2. Error terms are normally distributed with mean zero(not X, Y)
   3. Error terms are independent of each other
   4. Error terms have constant variance (homoscedasticity)

   Steps that we take while building a model:
   1. Data reading
   2. Data understanding and Exploratory Data Analysis to identify the unwanted features, probabilities of multicollinearity etc.

3. Data preparation involving conversion of features to its correct data types, converting data values to understandable format, the conversion of True/False, Yes/No columns to 1/0 values, converting the categorical variables using dummy encoding, taking care of the outliers etc
4. Train-Test data splitting
5. Rescaling the data, either using the minmax scaling or standardisation, based on the use case
6. Creating the target variable and independent variables from the test data
7. Feature selection – If there are a lot of features, feature selection can be done using Recursive Feature Selection else manually.
7. Building the linear model either by using sklearn or by statsmodel if detailed statistic summary is required
8. Adding/removing variables till the cycle where the model has all variables with p values less than 0.05, VIF less than 10, high r-squared, adjusted R squared and prob(F-statistics).
9. Residual analysis to check all the assumptions of Linear Model are followed by the model
10. Predicting the test data's dependent variable after proper scaling of the test data's independent variables
11. Evaluating the model for the test data results

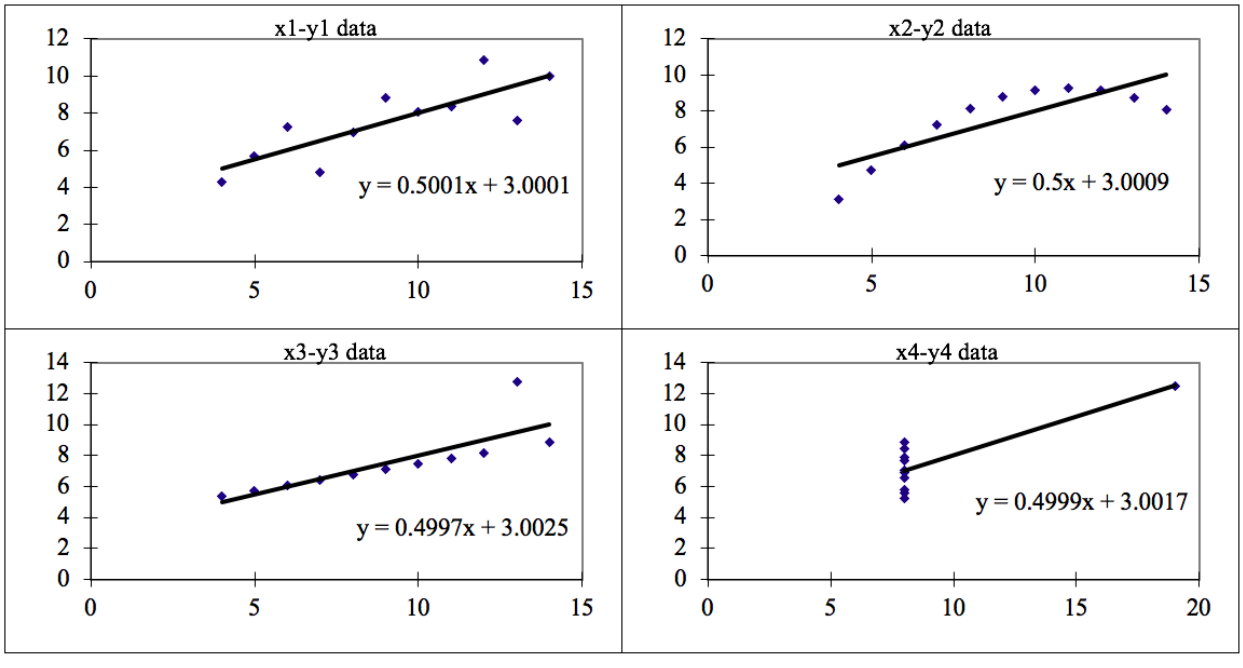
2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is defined as a group of four data sets that are nearly identical in simple descriptive statistics, which provides the same statistical information that involves variance, and mean of all x, y points in all four datasets, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It states the importance of visualizing the data before applying various algorithms to build models which shows that the data features must be plotted in order to see the distribution of the samples that can help to identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. The Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind

of datasets. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:

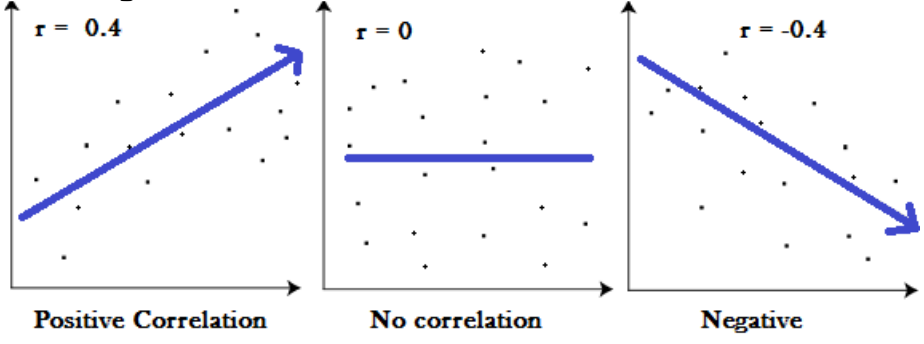


Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a roughly

linear relationship with some variance. Dataset II fits a neat curve but does not follow a linear relationship. Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well. We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

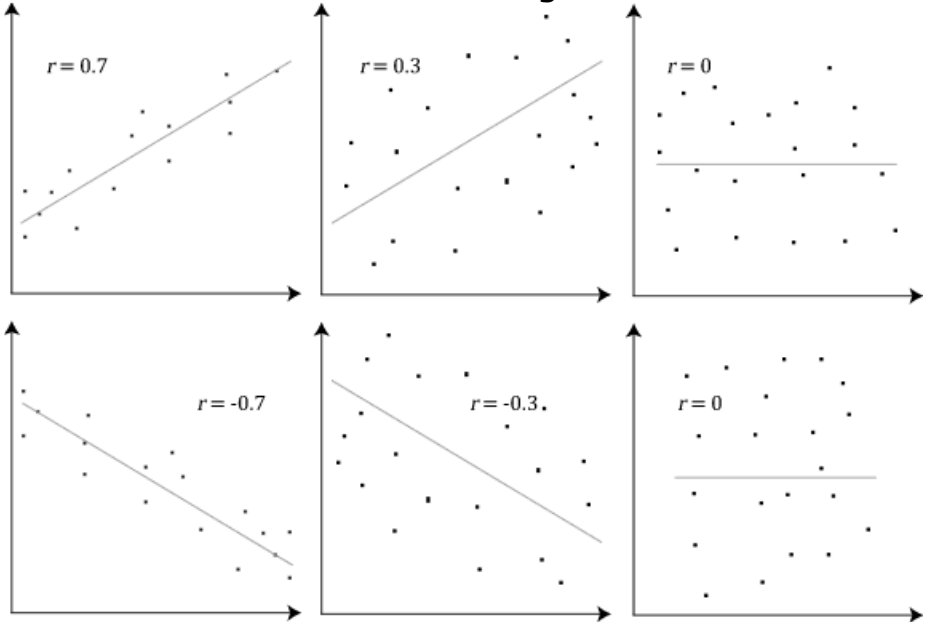
3. What is Pearson's R?

The Pearson product-moment correlation coefficient (Pearson correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r.
A Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are

included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



Pearson's r measures the degree of correlation or correlation coefficient between 2 numerical variables.
Its value varies between -1 and 1.
r = 1 means the data is perfectly linear with a positive slope (i.e both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
0<r<5 means there is a weak association
5< r< 8 means there is a moderate association
8 < r means there is a strong association

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

   Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on the same

scale in regression. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If we don't have comparable scales, then some of the coefficients obtained by fitting the regression model might be very large or very small as compared to the other coefficients. In regression, it is important to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means. 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. Machine Learning algorithm works on numbers, not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways:
1. Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
2. Standardization Scaling: Standardization 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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python

One disadvantage of normalization over standardization is that it loses 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?

   The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. n order to determine VIF, we fit a regression model between the independent variables.

It is calculated by taking the ratio of the variance of all beta values of a model divided by the variance of a single beta value if it were fit alone.
If there is a perfect correlation, then VIF = infinity which 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).

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

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. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It helps to identify whether the distributions are Gaussian, Uniform, Exponential or even Pareto.

   Few advantages:
   1. It can be used with sample sizes also
   2. 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 the following scenarios:
   If two data sets —

1. Have come from populations with a common distribution
2. Have a common location and scale
3. Have similar distributional shapes
4. Have similar tail behaviour

Interpretation:
Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to a straight line at an angle of 45 degrees from the x-axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degrees from the x-axis