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?

A. From the model created, following categorical variables are still retained after feature elimination and have significant impact –

a) The **weather situation** contributes to the sales only when it is case 4 -, i.e., **Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.** The coefficient obtained in this case is -0.2875, by which we can infer that the sales plummet by a factor of 0.2875 when the above conditions are active

b) **Weekdays** do not have a significant impact on the dependent variable, as the coefficients do not have a significant magnitude.

c) For the **months** categorical variable, we can find a significant increase in sales in the month of **September**, with the coefficient being 0.4203

d) Lastly, the **year** 2018 was not good for the bike sales as the coefficient observed is -0.2478

---

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

A. As we know, the categorical variables in linear regression have to be encoded with dummy variables to convert into numerical ones.

This is done by using one – hot encoding, where each state of the categorical variable is represented by a dummy variable with two states, on and off, represented by 0 and 1.

For example, the season variable in the assignment had 4 states – spring, summer, winter and fall.

The encoded dummy variables for this would be as follows:

	Spring	Summer	Winter	Fall
Spring	1	0	0	0
Summer	0	1	0	0
Winter	0	0	1	0
Fall	0	0	0	1

However, one observation can be made from the above table – if we treat fall as one state and the other 3 seasons as the second – it converts into a binary variable with only 2 states –

	Fall	Not Fall
Fall	1	0
Not Fall	0	1

Which means that the 4th variable is actually redundant, and the above information can be represented using only 3 variables – Spring, Summer and Winter, and when all of them are **off** or 0, it means the season is fall.

This means that any categorical variable with n states can be represented by (n-1) dummy variables.
This eliminates one feature from the final model for each categorical variable, which is very important as a higher number of redundant features will cause the final adjusted r squared for the model to reduce, as the model will be penalized one extra time for the nth state column we include.

---

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

A. Looking at the pair plots of all the variables with **cnt**, we can conclude that **temp** has the highest positive correlation with the target variable, as the relationship among them is almost linear.

---

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

A. The following methods were used to validate the assumptions:

   a) Seaborn pair plots to validate linear relationship of the variables with the target variable.

b) **Heatmap** using Pearson's coefficient and **Variance Inflation Factor** test after each iteration to determine collinearity of variables. Variables with high collinearity were detected and eliminated at each step.

c) Plotting the error terms against the mean of the dataset for homoscedasticity test. As the errors were equally randomly distributed across the mean, no visible relationship was found.

d) Histogram plot of the error variables in the train set to check for normal distribution. The error values were found to be almost normally distributed and centred around 0.

---

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

A. a) **Weather situation**, which has a significant negative impact on the final sales depending on the conditions.

b) **Windspeed,** which also has a negative impact on the sales, increased wind speed leads to reduced sales.

c) **Month** categorical variable, a significant increase is observed in the second half of the year, peaking in **September** with a positive effect on the bike sales.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A. Under the category of supervised learning in Machine Learning problems, where we are given a labelled data set to train our algorithm, there are two approaches to solve them –

a) Regression – Used to predict the relationship between the dependent variable and the predictor variables when the dependent variable is **continuous** in nature.

b) Classification – Used when the dependent or target variable is **categorical**, and the purpose is to segregate it into one of 'n' labels.

Regression itself has couple of algorithms, each attempting to solve the problem in a different manner.

The simplest of them is **Linear Regression**, in which we algorithm tries to create a **best fit line** that passes through maximum of the observed data points. This line is then extrapolated and used to predict the output variable values for unknown input values.

Here, the dependent variable is assumed to be y and the independent variable is assumed to be x.

Linear Regression itself takes two forms depending on the nature of the relationship between the variables –

1) **Simple Linear Regression** is used when the dependent or target variable has a relationship with only one independent variable.

The best fit line that then forms can be represented by:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where, β_0 is the intercept of the line, β_1 is the coefficient of the independent variable x and ϵ is the error margin.

2) **Multivariate Linear regression** is obtained by extending the concept of SLR to multiple independent variables that have a relationship with the target variable. Assuming the multiple independent variables to be x0, x1, x2 etc. we get –

$$y = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \epsilon$$

Getting the best fit line:

Here, as seen above, the values that determine the accuracy of our algorithm are β_0 and β_1 for the best fit line. They will predict the value of y for any unknown value of x.

To reach the best possible values for β_0 and β_1, we will have to go through a trial and error process. The aim of regression is to produce a best fit line such that the difference between the observed and predicted values, i.e., the **error** is minimum.

Now as we know, the error differences may be positive or negative, so to get an accurate representation of the error, we use various methods of calculating the difference, and the function which represents this calculation is termed as the **cost function**, usually denoted by J**.**

For linear regression, this cost function is given by a method called **Ordinary Least Squares**. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data points we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seek to minimize.

Mathematically, it can be represented by –

$$\boldsymbol{J} = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

And our goal is to minimize J, by changing the values of β_0 and β_1 on each iteration.

Gradient Descent:

Depending on the data points and number of independent variables, the algorithm might run very slow or miss the optimal values of β_0 and β_1 during training. To optimize this step, we follow a procedure called as **gradient descent**.

At each step of the iteration, we define a scaling factor α or θ, also known as the learning rate, which is the amount by which we change the values of β_0 and β_1. We calculate the local maxima and the global minima of the cost function **J**, and depending upon how far we are to the optimal values, we increase or decrease the learning rate.

Since this learning rate is independent of the iterations and is fixed for a given model, it is also termed as the **hyperparameter** of the model.

Assumptions

Before running linear regression on our data, there are some assumptions we make about the data set, and which have to hold true for the linear regression algorithm to produce a decent predictive model.

1. **Linearity** – We assume that the dependent and independent variables show a reasonably linear relationship, or else linear regression won't make sense. This can be verified by plotting them against each other and checking whether the data points make vaguely a straight line or not.

2. **Multicollinearity –** In case of multivariate linear regression, there should not be any relationship between the independent variables themselves. If one independent variable can explain another, then it would be impossible to determine which of them caused the output variable to change by what factor.

   This can be verified and enforced calculating Pearson's correlation coefficient, conduction variance inflation factor test, and other methods.

3. **Homoscedasticity –** We assume equal variances of all the observed data points form the mean, i.e., there is no observed trend between differences of the output data points. If this happens, then the target variable would be able to explain itself without the need of any input variable.

4. **Normal Distribution of Residuals –** The residuals derived from RMS are all normally distributed, and are centred around 0. This also applies to input data, and if the input is not normally distributed, we apply some transformations on it to convert it to a normal distribution.

---

2. Explain the Anscombe's quartet in detail.

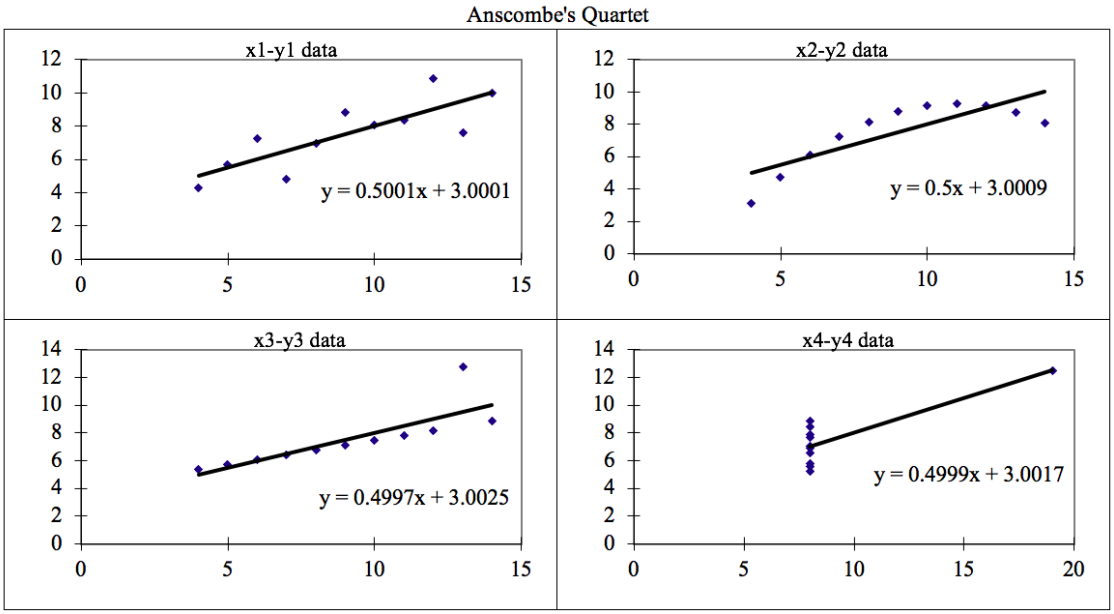
A. The Anscombe's quartet is a reminder of why statistical inferences are not always accurate, and why we need to perform data visualization and cleaning before attempting to model it.

According to Wikipedia, Francis Anscombe was a well-known statistician who happened to discover 4 datasets that, when modelled, all of them showed a linear relationship between the variables, but in reality, none of them were.

The datasets: (source: Google Images)

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Now, when a scatter plot is used to visualize all of these datasets, this is what it looks like:



We can get the following observations from the above image:

1. The first dataset does show a fairly linear relationship between x and y, and thus the best fit line makes sense.

2. The second dataset has a **non-linear** relationship between x and y, however, due to the nature and magnitude of the observed values, a linear relationship was inferred. This Is why one of the assumptions before going for linear regression is to ensure an **observed** linear relationship between the variables.

3. There is a clear outlier in the data set towards the end, which is not going to be predicted by the model, as the predicted value is nowhere near the actual value at that point. This gives importance of removing outliers from the dataset.

4. The fourth one should not have produced a linear model at all, however the one outlier at the end skewed the inferences and hence we get a best fit line nevertheless. This also goes towards outlier treatment.

Therefore, Anscombe's quartet helps us remember that it is easy for a linear model to mistakenly fit a line through any sort of data, so we must be careful to perform data cleaning beforehand.

3. What is Pearson's R?

A. The Pearson Product-Moment Correlation Coefficient (PPMCC) or Pearson's correlation coefficient represents how linearly related two variables are.

More specifically, if we get data points of two variables and attempt to draw a best fit line between them using linear regression, the Pearson's coefficient represents how far the actual data points are from this best fit line.

It is represented by r, and is given by the following ratio –

$$r_{xy} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Since it is a ratio, it will never exceed 1, however since variables may have a negative relationship as well, as in, one variable increase while the other decreases, it can go to -1 as well.

Therefore, the value of r ranges from -1 to +1.

A +1 indicates **perfect positive correlation**, which means that every observed value of y lies on the best fit line between x and y, and y increases as x increases.

A -1 indicates **perfect negative correlation**, which means that every observed value of y lies on the best fit line between x and y, and y decreases as x increases.

A 0 indicates no correlation, as y and x do not exhibit linear relationship at all.

The Pearson's coefficient can be used to determine the strength of the linear regression model, or to calculate the collinearity levels between two independent variables.

---

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

A. Scaling is the process of changing the range of an input variable by enlarging or diminishing it by a numeric quantity.

For example, if my input variable called **temp**, is from a range of 0-100 degrees Celsius, I can divide it by 100 to make it 0 to 1, multiply it by 5 to make it 0-500, and so on.

The mathematical operation performed on the variable is called the **scaling function**. There are lot of scaling functions used in machine learning.

Scaling is very important in linear regression as it helps the model in calculations – if all the variables are in the same range or a similar range of values, it becomes easier for the model to perform gradient descent and the learning rate is achieved way faster.

Otherwise, it may take up to a long time to determine different coefficients for differently ranged variables, and some coefficients may turn out to be in the range of 10000's whereas

others would be lesser than 1. This makes the output equation difficult to infer from, and may lead to incorrect interpretations.

Normalized vs Standardized Scaling –

Normalized Scaling, also known as min-max scaling, transforms all the variables into the same range of values, based on the minimum and maximum value of X. Mathematically, it is represented by –

$$X_{scaled} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

This scaling is usually done when the input variables are not in the same range of values, and are not already normally distributed. It always scales them between -1 and 1, which makes it easy for the model to quantify these values.

Standardized scaling, on the other hand, assumes an already gaussian distribution of the input variables, and attempts to scale them by using their mean and standard deviation. It aims for **zero** mean and a standard deviation of **1**.

Mathematically,

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

As we know, this is also the formula to get the Z-score, so this method is also called as Z-score scaling. Its not bounded to a specific range, but preserves the shape of the distribution.

---

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

A. As we know, the formula of VIF is given by:

$$VIF = \frac{1}{1 - R^2}$$

Where, R is the Pearson's correlation coefficient.

As we know, the value of R ranges from -1 to +1, -1 indicating a **perfect negative correlation** and +1 indicating a **perfect positive correlation**.

In both these cases, form the above formula, VIF becomes **infinity,** therefore indicating that there is a perfect positive correlation between the two variables. This is a sign of collinearity and one or both of them have to be removed from the model.

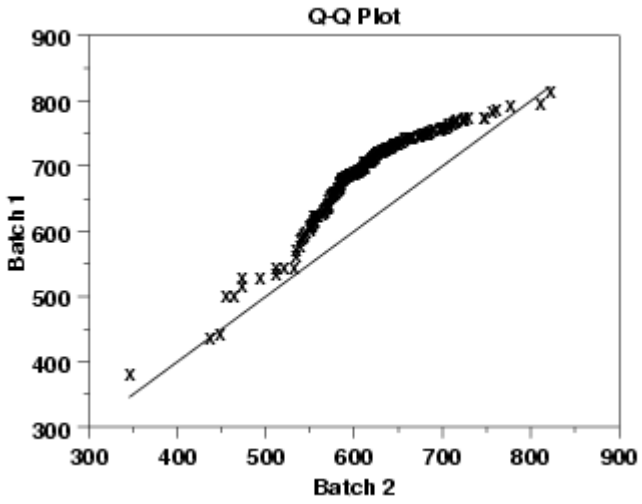
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A. A Q-Q plot, also know as a quantile-quantile plot, is used to determine whether 2 datasets come from the same sample distribution or not.

It is essentially a distribution of the **quantiles** of the dataset, which are indicators of how much of the data lies below that quantile, for e.g., a 0.25 quantile corresponds to 30% of the population, 0.5 corresponds to 50% and so on.

There is a 45-degree slope line of reference provided – if both of the datasets originate from the same source, all the points on the Q-Q plot will lie on this line, otherwise not.

Example –



In this case, not many of the quantiles lie on the line, hence it can be inferred that these two batches are not from the same original dataset.

In linear regression, this can be used to infer whether the residuals follow the same normal distribution as the input data set or not – this is important as similarity between theoretical quantiles and the observed residuals will verify the significance of the linear model, and whether it accurately depicts the dataset or not.