Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3marks)

A1. Following are my inferences about the effect of categorical variables on the dependant variable:

- The first categorical variable with the most effect on the dependent variable count (‘cnt’) is **light rain or snowy weather** (specifically described as Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).

  Among the categorical variables, this variable has the most effect on count, but it has a **negative** effect. Count is expected to **reduce** with unit increase in 'light rain / snow'.

- The next categorical variable with most effect on count is **year** (‘yr). It has a **positive** effect on count. This means that count is expected to **increase** with unit increase in 'yr'

- **Holiday** the next highly affecting categorical variable. It has a **negative** effect on count. Count is expected to **reduce** with unit increase in 'holiday'.

- **Winter** season has the next highest effect on count. It has a **positive** effect on count. So, count is expected to **increase** with unit increase in 'winter'.

- **September** month has the next highest effect on count, count is expected to increase with unit increase in 'sep'

- **Misty / cloudy** weather has the next highest effect on count. It has a **negative** effect on count. So, count is expected to **reduce** with unit increase in 'misty / cloudy'

- **Summer** season has the next highest effect on count. It has a **positive** effect on count. Count is expected to **increase** with unit increase in 'summer'

- **Spring** season has the next highest effect on count. It has a **negative** effect on count. So, count is expected to **reduce** with unit increase in 'spring'

Q2. Why is it important to use ‘drop_first = True’ during dummy variable creation? (2mark)

A2. Using ‘drop_first = True’ is also referred to as ‘dummy encoding’. It is important to use dummy encoding because it removes the extra column created during dummy variable creation (pd.get_dummies function). This extra column that is removed was a redundant column. So by using dummy encoding we reduce the correlations among dummy variables.

I’ll explain this further with the help of an example**.** If we have 3 types of values in a categorical column (eg: red, blue, green) and we want to create dummy variables for that column. If one record is not red or blue, then it is obvious that it would be green. So we do not need a third variable to identify the unfurnished records.

Before applying pd.get_dummies(drop_first = True)

Value	Indicator Variable		
Colour	Red	Blue	Green
Red	1	0	0

Blue	0	1	0
Green	0	0	1

After applying pd.get_dummies(drop_first = True)

Value	Indicator Variable	
Colour	Red	Blue
Red	1	0
Blue	0	1
Green	0	0

- Here 1,0 depicts Red
- 0,1 depicts Blue
- And 0,0 depicts Green

Hence, if we have a categorical variable with n-levels, then we need to create n-1 columns to represent the dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

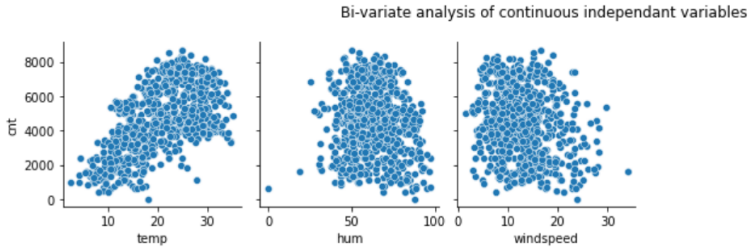
A3. Looking at the pair-plot of numerical variables plotted against the target variable 'cnt', the highest correlation can be seen in the scatter plot of **temperature** ('temp') on the X-axis and the dependant variable count ('cnt') on the Y-axis.

The correlation appeared to be positive, which means that increase in temperature would cause an increase in count. On plotting a heatmap, I discovered that the correlation among 'temp' and 'cnt' is indeed positive and it is 0.63, i.e., 63%.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A4.

- **Linearity -**
  The first assumption of Linear regression is that it needs the relationship between the independent and dependent variables to be linear. I checked this when I plotted the pairplot of numerical variables ('temp', 'hum', 'windspeed') against the target variable 'cnt' and saw that the scatter plots indicated somewhat linear relationships.



- **Mean of Residuals -**

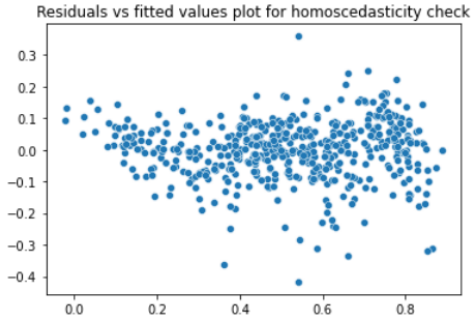
  Residuals as differences between the true value and the predicted value of the target variable. One of the assumptions of linear regression is that the mean of the residuals should be zero.

  When I tested this assumption, I found the mean of residuals to be extremely close to 0.

```
Mean of Residuals:  -2.27437894226029e-15
```

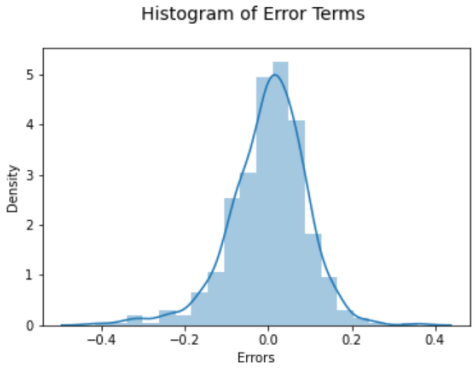
- **Homoscedasticity -**

  This assumption says that the residuals have equal or almost equal variance across the regression line. By plotting the error terms against y_train_pred, I checked that the assumption was being honoured.¶



- **Check for Normality of error terms/residuals -**

  To check if the error terms are normally distributed I had plotted the histogram of the error terms / residuals. The distributions of residuals/error terms was supposed to be centered around 0 and it was supposed to be resembling a normal distribution, and qualitatively looking at it; it indeed seemed so.



- **Interdependence of residuals -**

  When the residuals are autocorrelated, it means that the current value is dependent on the previous (historic) values and that there is a definite unexplained pattern in the Y variable that shows up in the error terms. We want there to be no autocorrelation of residuals. In our dataset the residuals were independent of each other.

- **No Multicollinearity -**

  In regression, multicollinearity refers to the extent to which independent variables are correlated. I ensured that there is no multicollinearity in my model by making sure that the VIF of all my predictor variables is below 5.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A5. My final model ('lr_8') indicated the top 3 features contributing significantly towards explaining the demand of the shared bikes as the following:

1. **Temperature** has the maximum effect on count. The effect is positive, represented by the coefficient 0.478. This means that count is expected to increase most noticeably with unit increase in 'temp'.

2. **Light rain / snowy weather** has the second highest effect on count. The effect is negative, represented by the coefficient -0.285. So, count is expected to reduce with unit increase in 'light rain / snow'.

3. The variable **year** has the next highest effect on count. It has a positive effect, represented by the coefficient 0.234. So, count is expected to increase with unit increase in 'yr'.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4marks)

A1. Linear Regression is based on supervised learning. Which means that there is a dependent variable (here, a continuous variable) to be predicted based on independent variables. It performs a regression task, and is used for finding out the relationship between variables and for forecasting. It can be of two types, simple linear regression and multiple linear regression. Following are the steps involved:

Reading, understanding, visualising the data

- This provides an intuitive understanding of the dataset. It involves data cleaning - viewing head, data types, statistical summary, checking for null values
- Next, we perform data manipulation - eg: ensuring that all categorical variables have string values and not numerical.
- We also perform data visualisation and exploratory data analysis in this step.

Data preparation

- This is when we create dummy variables on all categorical variables as the model needs only continuous/numerical values. We also convert binary variables to 1 and 0.
- Next we do a training set and test set split, often as a 70% and 30% split. Train set will be used to build the model and use the model to test and predict the outcome.
- We also perform rescaling of features in this step - we can go for either min-max scaling or for or standardization so as to convert variables into a comparable scale.
- Finally we divide into X and Y sets for the model building.

Training the model

- This is when learning the coefficients and intercept takes place using the OLS (ordinary least squares) method
- Next we select the right variables by making sure to drop insignificant predictors (p-values > 0.05)
- Finally we drop variables one by one having high multicollinearity; those with VIF > 5

Residual analysis

- Here we perform residual analysis on the train data i.e, check if the error terms are normally distributed and are centred around zero; additionally we do other checks of ensuring that all linear regression assumptions are being honoured.

Prediction and Evaluation of Model

- Since we have our final model now, we make predictions using the final model on the test dataset.
- Finally we evaluate the model by calculating the parameters of R-squared, Adjusted R-squared and RMSE on the test set. We can compare these parameters by calculating them on the training set to ensure that we are satisfied with our final model - having a good enough R-squared and not much gap in between the R-squared of test and train sets (to check for overfitting)

Q2. Explain the Anscombe's quartet in detail. (3marks)

A2. Anscombe's quartet is a group of 4 datasets that have nearly identical statistical properties, yet are very different when they get graphed. Each of the dataset consists of eleven points. These datasets were constructed in 1973 by the statistician Francis Anscombe to show the importance of graphing data before analyzing it and the effect of outliers on statistical properties of data.

Following image shows the datasets:

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

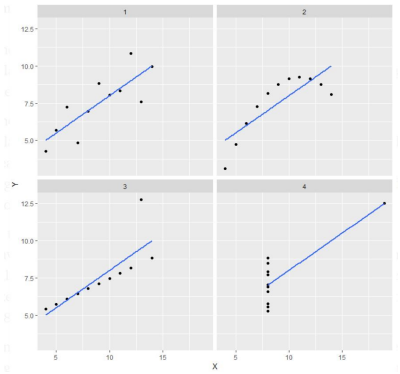
When we calculate the statistical summary of these datasets we get extremely similar results:

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

- The average *x* value is 9 for each dataset
- The average *y* value is 7.50 for each dataset
- The variance for *x* is 11 and the variance for *y* is 4.12
- The correlation between *x* and *y* is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation $y = 0.5x + 3$

When we plot the scatter plots of these datasets we see that although the statistical information was giving us a different impression but when we plot the scatter plots, only then we get to know the actual composition of the datasets:



Interpretation of scatter plots:

- On the top left we see that there seems to be a somewhat linear relationship between x and y.
- Top right shows a non-linear relationship between x and y, more like y increasing with x upto a point and then reducing.
- Bottom left has a perfect linear relationship for all the data points except one which looks like an outlier.
- Bottom right graph shows an example when one high-leverage point is enough to produce a high correlation coefficient.

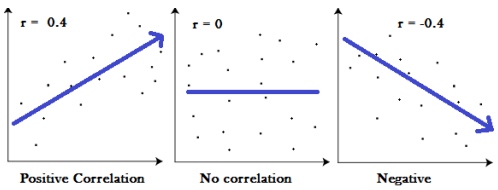
Conclusion

- The quartet is often used to prove the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship.
- This isn't to say that summary statistics are useless. They're just misleading alone by themselves. It's important to use them as just one tool in a larger data analysis process.
- Visualizing our data allows us to revisit our summary statistics and recontextualize them as needed.
- For example, Dataset II from Anscombe's Quartet demonstrates a strong relationship between *x* and *y*, it just doesn't appear to be linear. So a linear regression would be the wrong tool to use there, and we can try other regressions. Eventually, we'll be able to revise this into a model that does a great job of describing our data, and has a high degree of predictive power for future observations.

Q3. What is Pearson's R? (3marks)

A3. Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes and foot length.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank and speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-0.75| = 0.75, which has a stronger relationship than 0.65.

Pearson R / Pearson Product Moment Correlation (PPMC) shows the linear relationship between two sets of data. But it is not able to tell the difference between dependent variables and independent variables. For example, if we are trying to find the correlation between foot size and shoe size, we might find a high correlation of 0.8. But we get the same result with the variables interchanged. As in, that large shoe size causes a large foot size, which makes no sense. Therefore, we have to be aware of the data we are plugging in.

Pearson's coefficient also does not give us any information about the slope of the line; it only tells us whether there is a relationship.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3marks)

A4. Meaning of scaling:
Scaling refers to putting the values in the same range or same scale so that no variable is dominated by the other. It is performed during the data pre-processing.

For example, if an algorithm is not using scaling then it will consider the value 3000 meters to be greater than 5 kilometers; and so the algorithm will give wrong predictions. So, we use scaling to bring all values to the same magnitude and tackle this issue.

Scaling can vary results of certain machine learning techniques but have a minimal or no effect in others. Techniques like linear regression, logistic regression, neural network, etc. that use gradient descent require the data to be scaled.

Reason scaling is performed:
Formula for gradient descent:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature.

To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Difference between normalized scaling and standardized scaling:

- Min-Max Scaling: It re-scales a feature or observation value with distribution value between 0 and 1. Following formula is used to perform it:

  $$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- Standardized Scaling: It e-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. Following formula is used to perform it:

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3marks)

A5. If there is perfect correlation, then VIF = infinity; i.e., completely redundant variable. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) to become infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

In order to understand what 'perfect multicollinearity' means we can understand what VIF is and how it is calculated behind the scenes: VIF is an index that provides a measure of how much the variance of a regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

For example, we would fit the following models to estimate the coefficient of determination R_1 and use this value to estimate the VIF:

$X\_1=C+ \alpha\_2 X\_2+\alpha\_3 X\_3+\cdots$
$VIF\_1=1/(1-R\_1^2)$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$X\_2=C+ \alpha\_1 X\_1+\alpha\_3 X\_3+\cdots$
$VIF\_2=1/(1-R\_2^2)$

- A large value of VIF indicates that there is a correlation between the variables.
- If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
- This would mean that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).
- 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.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3marks)

A6. Quantile-Quantile (Q-Q) plot, is a plot that helps us determine if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It also helps to determine if two data sets come from populations with a common distribution.

Importance of a Q-Q plot in linear regression is that 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. Advantages are that 1. It can be used with sample sizes also and 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 following scenarios:

- If two data sets come from populations with a common distribution
- If two data sets have common location and scale
- If two data sets have similar distributional shapes
- If two data sets have similar tail behavior