Assignment-based Subjective Questions

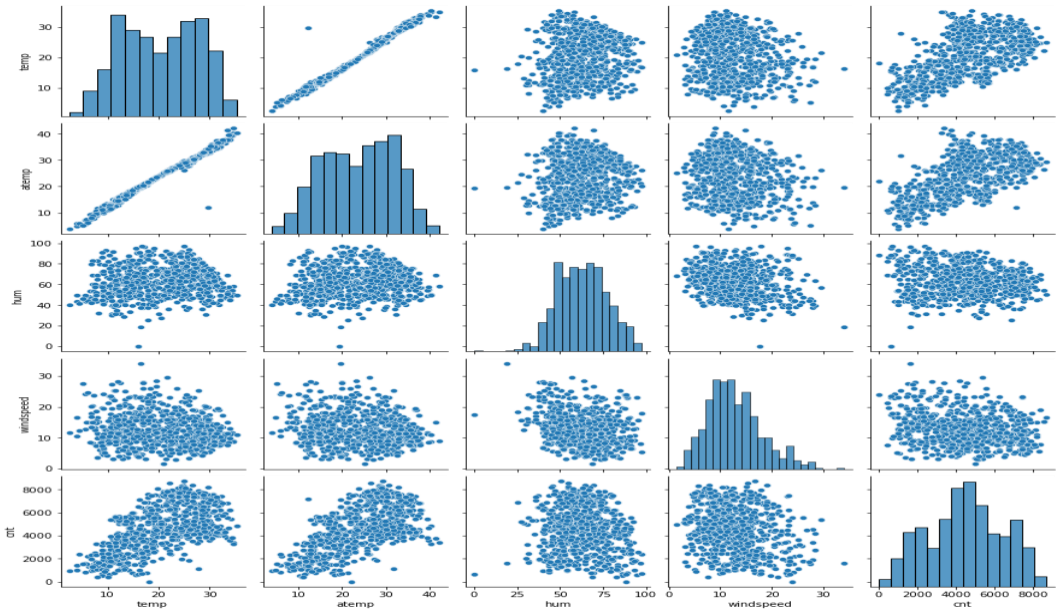
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   The categorical variable in the dataset are **season, mnth, holiday, weekday, weathersit, yr.** These are visualized using the boxplot. These variables have the following effect on dependent variable:-
   1. Season – The boxplot showed that spring season has lowest demand while fall sees the highest
   2. Mnth – September saw the highest number of rentals while January saw the lowest
   3. Holiday – rentals reduced during holidays
   4. Weekday – Sunday and Thursday have higher variations in rental than any other day
   5. Weathersit –demand seems to increasing when weather is clear
   6. Yr – demand for next year 2019 is high.

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

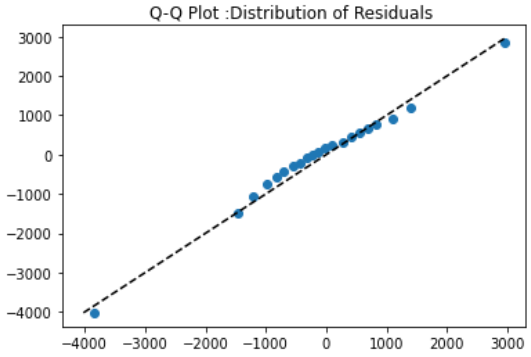
   It is important to use drop_first, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables decreasing multicollinearity.

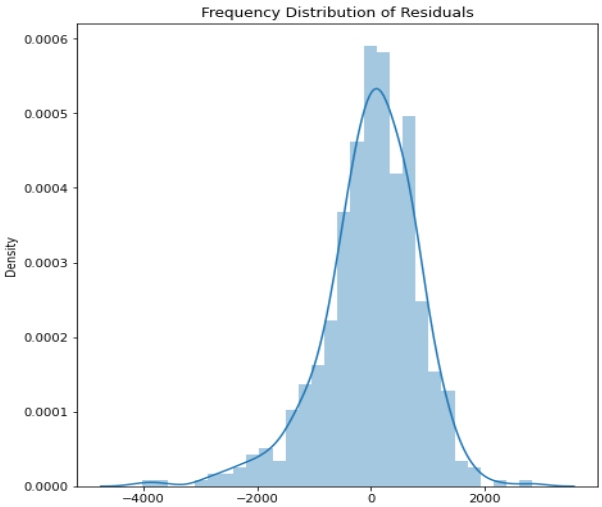
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

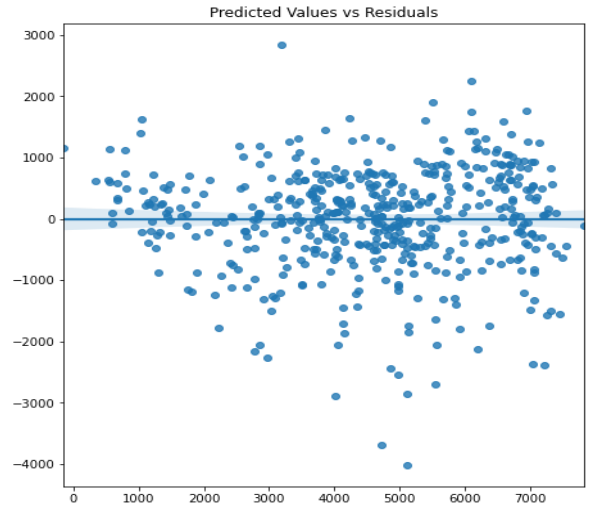


'temp' and 'atemp' seems to have highest correlation with target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   -We plotted two graphs to check for normality of error distribution and its mean around 0, as it is one of the assumptions of Linear regression after building the model.
   -We also plotted QQPlot to confirm if the errors are normally distributed.
   -We also plotted regplot or scatter plot to see the any pattern in errors between errors and







   We also plotted regplot or scatter plot to see the any pattern in errors between errors and Y test data.

Hence we can confirm that

- Residual errors follow a normal distribution with mean=0
- Variance of Errors doesn't follow any trends
- Residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend.

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

```
1.temp              3738.377584
2.yr                2038.011476
3.Weathersit-spring    -1262.612994
```

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

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. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Regression is most commonly used predictive analysis model,
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
Linear regression is based on popular equation "**y=mx+c**".
In regression we calculate the best fit line which describes the relationship between dependent and independent variables. Regression is performed when the dependent variable is of continuous type and predictor variables could be of any data types.
Regression is broadly divided into simple linear regression and multiple linear regression

1. **Simple Linear Regression: SLR** is used when dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression: MLR** is used when dependent variable is predicted using only multiple independent variable.

Equation for of SLR and MLR is :

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\epsilon$$

Y= Dependent variable

β_0=intercept (constant term)

β_1=coefficient of X1 independent variable

β_2=coefficient of X2 independent variable

X=predictor or independent variable

ϵ=Errors

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

DATASET

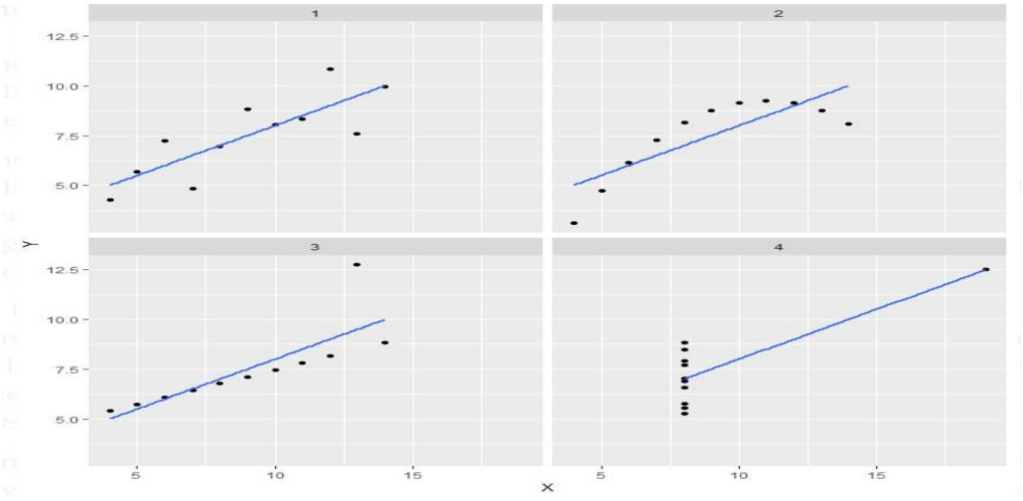
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

STATISTICAL SUMMARY OF DATASET

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties**, yet appear very different when graphed.**



Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

Pearson's R is a numerical summary of the strength of the linear relationship association between variables. It is also known as Pearson correlation coefficient
Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.
Pearson correlation coefficient is a measure of the strength of a linear association
In simple terms it tell us *can we draw a line graph to represent the data*
A value of ± 1 indicates a perfect degree of association between the two variables.
A value of 0 represents no linear association.

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

Feature scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Why?
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. *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.*

- **Normalization/Min-Max Scaling:** Normalization is generally used when you that the distribution of your data doesn't follow Gaussian distribution, brings all of the data in the range of 0 and 1.

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

- **Standardization Scaling:** Standardization on other hand could be helpful in cases where data follows a Gaussian distribution. However, this doesn't have to be necessarily true, also unlike normalization does not have a bounding range

  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**).

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

*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? (3 marks)

If there is perfect correlation, then VIF = infinity. This 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).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

It helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression 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 also it is used to show weather the error or residual have normal distribution hence confirming assumptions.

It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution

ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Possible interpretations for two data sets.

*a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.*
*d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.*