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:

Dependent variable: count – Total number of bike rentals

- Count increased in the year 2019 compared to 2018
- Clear days in fall had highest number of counts
- Count was highest in Fall across different weathers
- If weather is Cloudy/light-snow, people will still go out in bikes in Fall/winter, but not so much in summer/spring
- May to October has pick counts, corroborates the findings from analysis of counts vs season
- Non-working day in the middle of the week sees highest counts

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

Ans:

If there are N levels of values, get_dummies will create N dummy columns by default.

N levels of values in a categorical variable can be explained by (N-1) dummy variables. For example – if it's not summer or spring or fall, then it must be winter. Hence it is important to drop the first column after dummy variable creation to get rid of one redundant column whose values can be well explained by the others.

But it is not mandatory to always drop the first one. As per business need, any one of the N dummy columns can be dropped, if it makes more sense in explaining the derived model to end users.

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

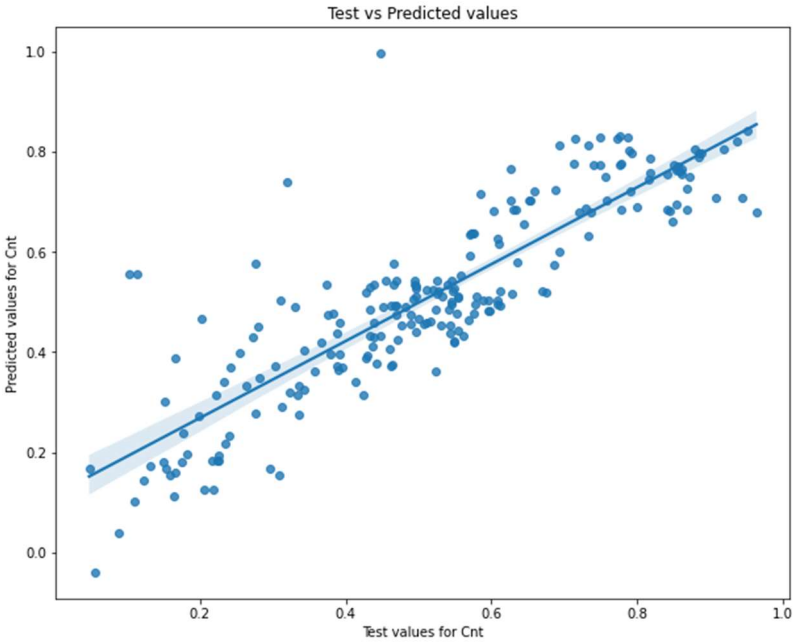
Ans:

Registered has the highest correlation (0.95) with target variable cnt.

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

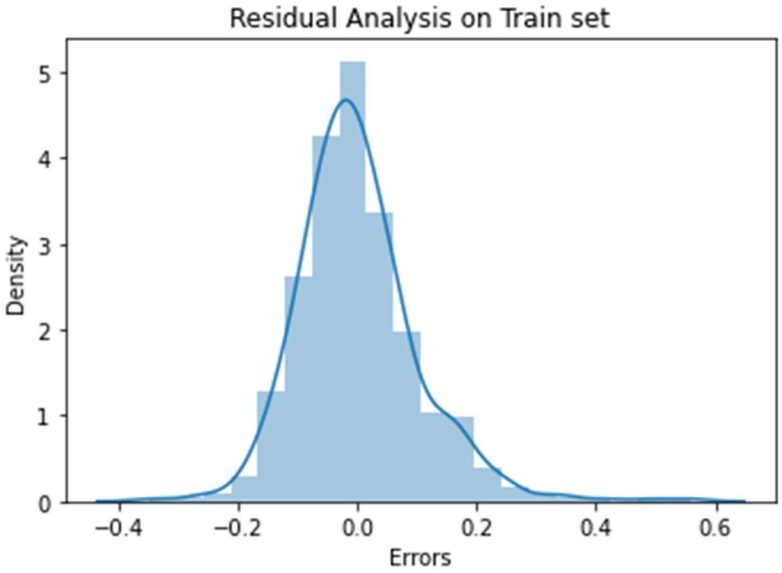
Ans:
Assumptions of Linear regression:

- **Linear relationship**: Assumption - There exists a linear relationship between the independent variable, x, and the dependent variable, y.
  For simple linear regression, it should be easy to test by plotting dependent vs independent variable. But for multiple linear regression, since there are multiple independent variable, we can test it by plotting test values vs predicted values.



As we can see, from the above plot, it shows a linear relation between tested and predicted values.
Linear equation obtained from final model:
cnt = 0.2481 + 0.2361 * yr - 0.0832 * holiday + 0.3592 * temp - 0.1567 * windspeed - 0.1534 * Spring + 0.0768 * Clear - 0.1974 * Light_snow_rain

- **Normality**: Assumption - The residuals of the model are normally distributed.

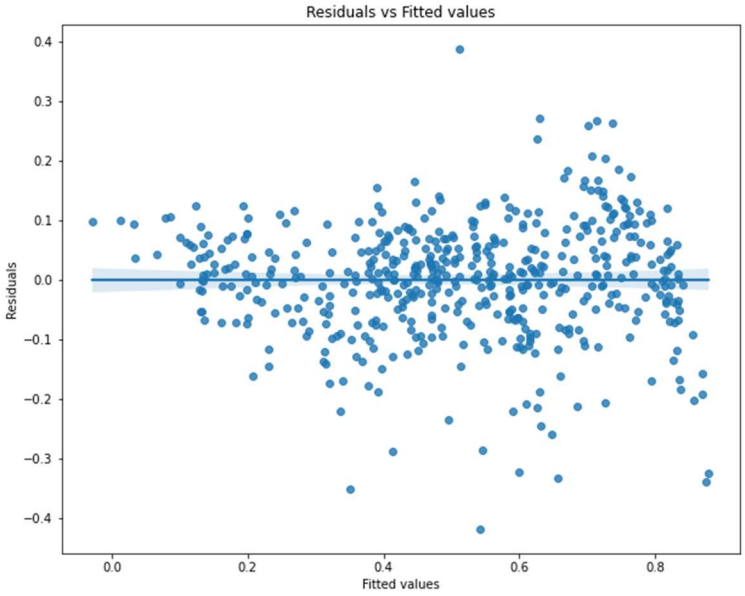


As we can see from the fitted model, residuals are very close to normally distributed with a slightly long tail.

- **Independence**: Assumption - The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
  This can be formally tested using Durbin-Watson test.
  As per Durbin-Watson test:
  - A test statistic of 2 indicates no serial correlation.
  - The closer the test statistics is to 0, the more evidence of positive serial correlation.
  - The closer the test statistics is to 4, the more evidence of negative serial correlation.

  As Durbin-Watson test result obtained in this case is 1.9914310225996599, which is close to 2, we can conclude that the residuals have no serial correlation

- **Homoscedasticity**: Assumption - The residuals have constant variance at every level of x.
  We can check this assumption by plotting residuals against fitted values as shown in the following figure.



As we don't see any obvious pattern, we can conclude assumption of homoscedascity is satisfied.

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:

Equation obtained from final model:

cnt = 0.2481 + 0.2361*yr - 0.0832*holiday + 0.3592*temp - 0.1567*windspeed - 0.1534*Spring + 0.0768*Clear - 0.1974*Light_snow_rain

Top 3 contributing features are temp, yr and Light_snow_rain.

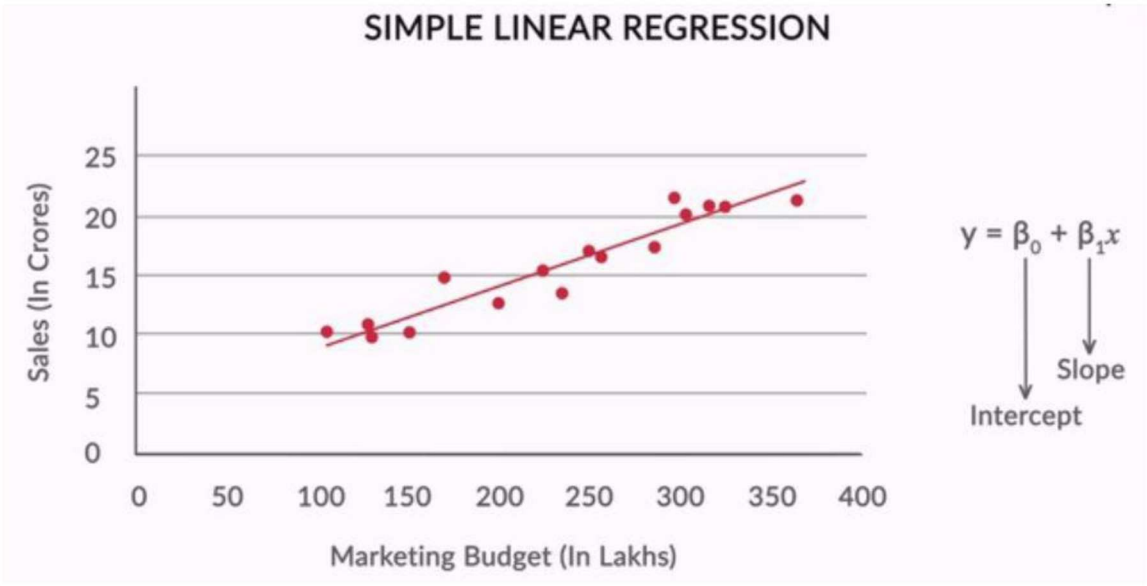
General Subjective Questions

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

Ans:

Linear regression is an approach for predicting the relationship between a dependent variable and one or more independent variables or predictors. When there is only one predictor, it is called simple linear regression. In case there are multiple predictors, it is called multiple linear regression.

In simple terms – linear regression draws a line or curve through all the predictor variable data points in such a way so that the vertical distance between the data points and the line is minimum.



Simple linear regression formula:

$$y = a_0 + a_1 x$$

Multiple linear regression formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Y: Dependent variable

X, X1…Xp: Independent variables or predictors

A0, b0: intercept of the line

A1, b1, b2…: coefficients of independent variables

ϵ : Error term

OLS or Ordinary least squares is one of the methods used in linear regression where parameters of a linear function are chosen by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable. But there are other methods also, for example WLS or Weighted least squares.

Cost function: Cost function helps to predict best possible values of the coefficients of the predictor, hence finding the best fitted line. In Linear regression, Mean squared Error (MSE) is used to calculate cost function. MSE calculates the average of squared error between predicted values and actual values.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Objective of linear regression model is to minimize this cost function. This is done using **Gradient descent**, where coefficients are randomly selected and iteratively updated to reach the minimum cost function.

Assumptions of Linear regression:

- **Linear relationship**: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
- **Normality**: The residuals of the model are normally distributed.
- **Independence**: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
- **Homoscedasticity**: The residuals have constant variance at every level of x.

References:

https://en.wikipedia.org/wiki/Linear_regression
https://en.wikipedia.org/wiki/Ordinary_least_squares
https://en.wikipedia.org/wiki/Linear_least_squares
Lecture notes

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

Ans: Anscombe's quartet consists of four different data sets that all have very similar summary statistics, but looks completely different from each other when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

The data looks like as follows:

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

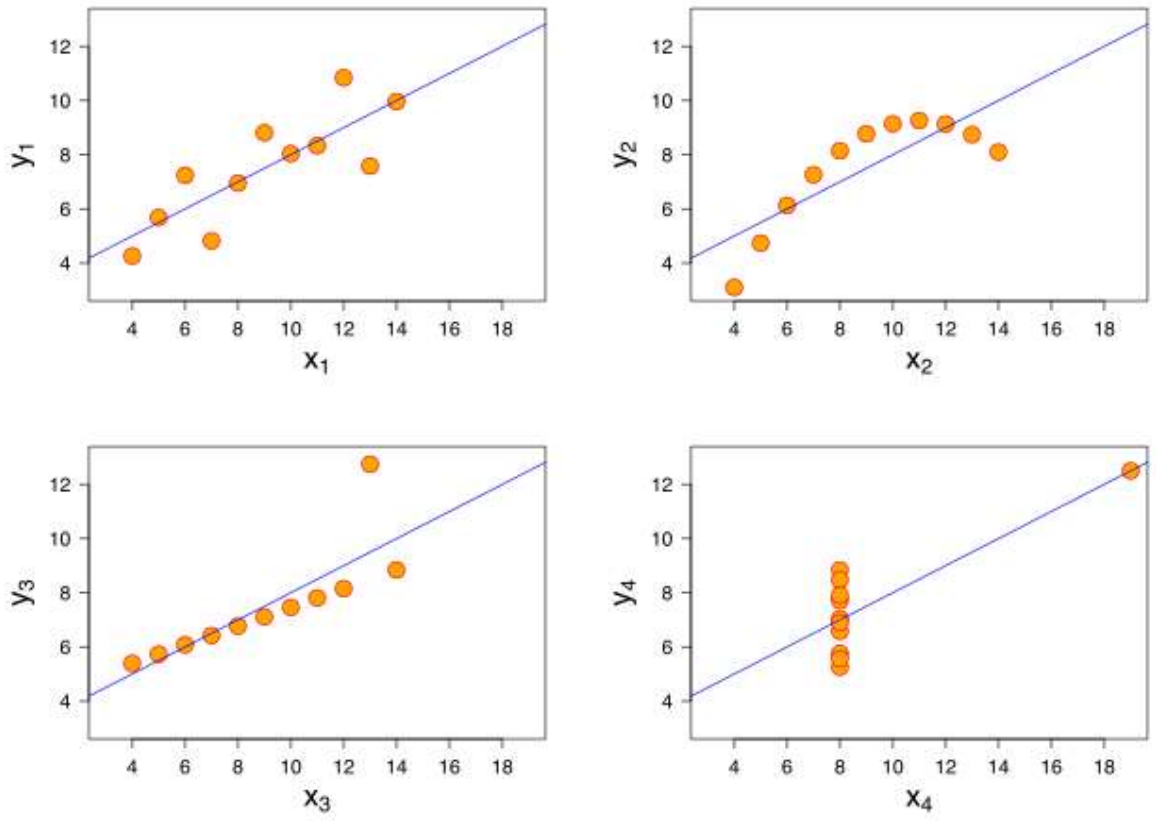
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary statistics for all four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively

Coefficient of determination of the linear regression	0.67	to 2 decimal places

Graphs of the four data sets:



- The first scatter plot (top left) appears to be a simple linear relationship
- The second graph (top right) is not linear.
- In the third graph (bottom left), the relationship is linear, but should have a different regression line. It is biased by the one outlier.
- Finally, 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 are not related.

Reference:
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

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

Ans: Person correlation coefficient or Pearson's R or more commonly known as just correlation coefficient is a measure of linear correlation between two sets of data. The measure can only reflect a linear correlation between two variables, and ignores many different types of relations. It has a value between -1 and +1. Correlations equal to -1 or +1 signifies all data points lying exactly on the line.

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

The sign of the correlation is determined by the regression slope – positive signifies a positive slope where y increases for increasing x. Negative signifies a negative slope, where y decreases with increasing x.

Reference:

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#:~:text=In%20statistics%2C%20the%20Pearson%20correlation,between%20two%20sets%20of%20data

Lecture notes

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

Ans:

Scaling – In machine learning, scaling refers to transforming values of predictor variables to the same range.

Why Scaling is important - Scaling is important as most of the times data sets consists of features with different magnitudes and units of values. For example: consider a data set consisting of two columns, shoe size and weight. shoe sizes may have values in the range 3 to 12 (UK size), but weight may have values in the range of 40 to 120 (Kg). In absence of scaling, machine learning model can become biased towards certain features because of their magnitude of values, not taking into account the unit in which those values are represented. Hence, we need to scale and bring the values to the same range before building models.

Difference between normalization and standardization - In normalization, scaling is done using min and max values of the feature. Feature values are mapped into the [0, 1] range.

Formula for normalization:

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

In standardization, we don't enforce the data into a definite range. Instead, we transform using mean and standard deviation. So, standardization also centralizes the data around mean 0. Standardization formula:

$$z = \frac{x - \mu}{\sigma}$$

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

(3 marks)

Ans: Very large values of VIF signifies presence of correlation between variables. So, an infinite VIF would mean perfect correlation between variables. It indicates that the corresponding variable can be exactly expressed as a linear combination of the other variables.
In this case we get R2 =1, hence 1/(1-R2) is infinity.

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

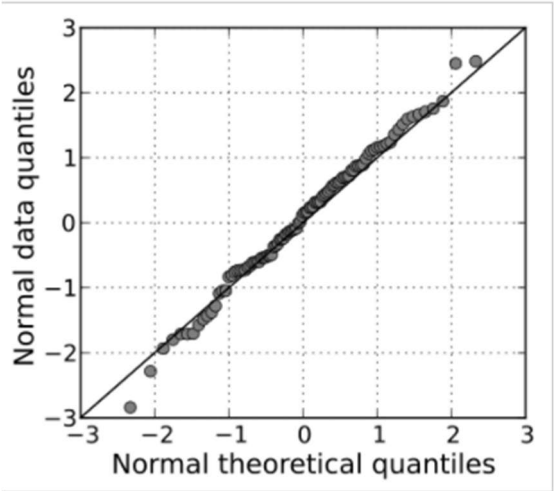
Ans: A Q-Q (quantile-quantile) plot is a graphical tool to help assess if two sets of data have similar theoretical probability distribution.

It is used to check following scenarios:

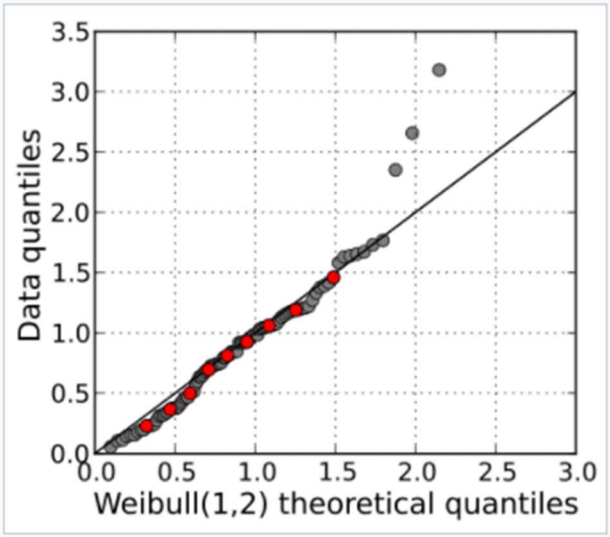
If two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

If the two distributions compared are similar, all data points in Q-Q will lie on the x=y line.



If the data sets are related, points will lie approximately on the x=y line.



A Q–Q plot of a sample of data versus a Weibull distribution. The deciles of the distributions are shown in red. Three outliers are evident at the high end of the range. Otherwise, the data fit the Weibull(1,2) model well.

Reference: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot