# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

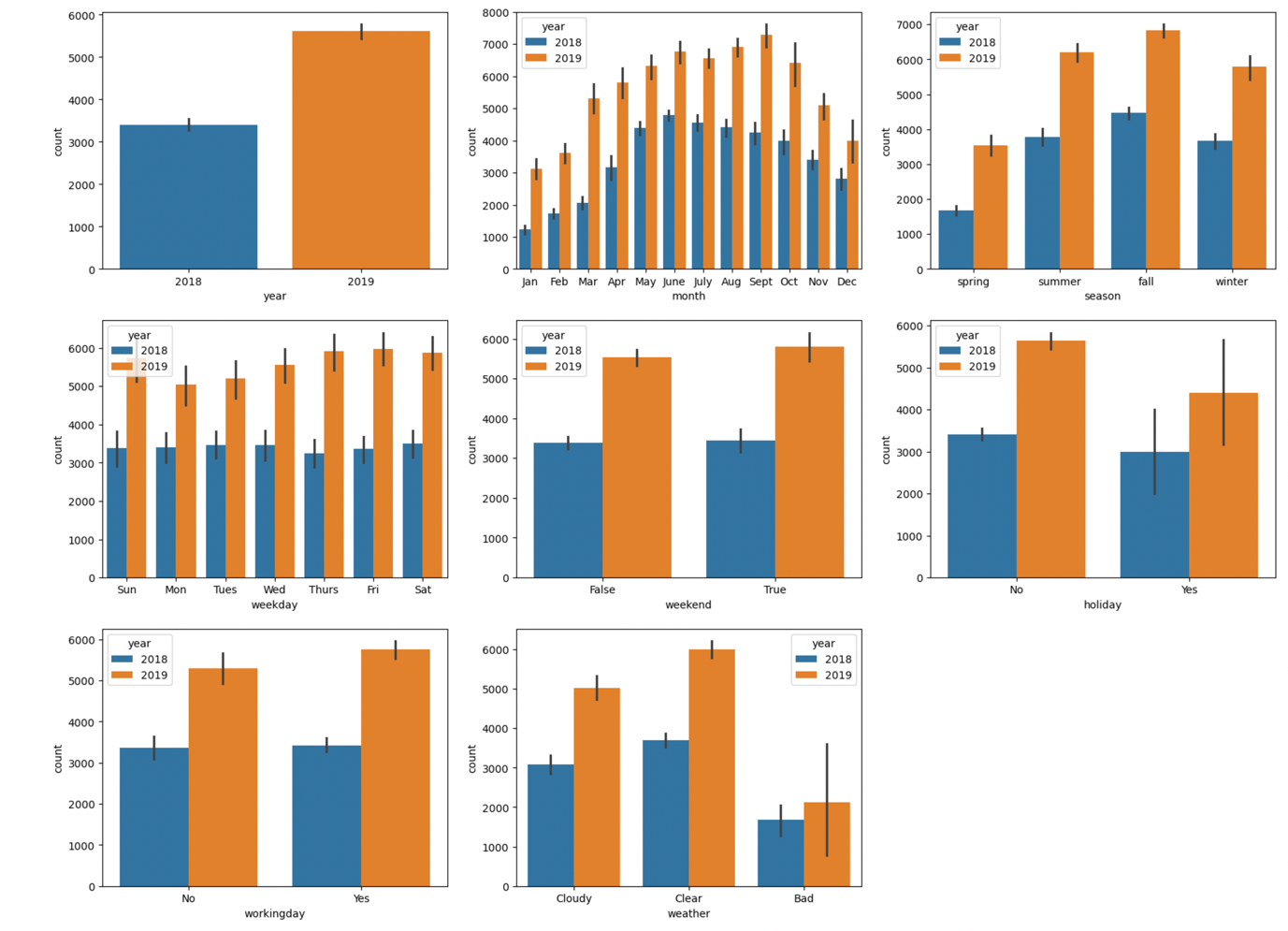
# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

* **Year**: Year 2019 has higher bookings compared to year 2018, means the bookings are increasing year on year and company can predict higher booking for upcoming year.
* **Month**: The month September & October should be considered by the company as they have a higher demand as compared to other months and December, January, February and July months have less bookings.
* **Weekday**: The count of rentals is almost even throughout the week and Monday seems to be the minimum bookings in the week.
* **Holiday**: Holiday time seems to be least demand for bookings.
* **Workingday**: Working day has slightly higher demand that non-working days ( i.e holiday or weekend).
* **Season**: The demand increases during summer and goes to peak during fall season and then starts decreasing during winter and least during Spring season, this same curve we can notice for Month variable as well, where the demand curve increases March onwards (i.e Summer season) then reaches peak during September (i.e Fall season), then starts decreasing during December (i.e Winter season) and is least during Jan and Feb(i.e Spring season)
* **Weather**: Demand is higher for clear weather and then for cloudy weather, and very less bookings for light rain/snow weather and absolutely no booking for severe/high rain/snow weather condition. That means weather a an important driving factor.

A group of blue boxes with numbers

Description automatically generated



# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

# Dummy variable creation is binary representation (0, 1) of categorical variables used in data preparation for machine learning model.Drop\_first=True is important to use during dummy creation because it reduce the extra column created during dummy variable creation. Hence it reduces the correlations and redundancy.

# And n-1 dummy variables will be able to predict the value of the nth dummy variable in a linear regression because the intercept will deal with the nth variable, So we do not need the nth dummy variable.

# Example: Suppose we have 3(n) variables: Furnished, Semi-furnished and un-furnished. We can only take 2(n-1) dummy variables, 0 and 1. So 1-0 will represent furnished, 0-1 will represent semi-furnished and 0-0 will represent un-furnished.

# # creating dummy variables for categorical features.

# day\_df = pd.get\_dummies(data=day\_df,columns=cat\_cols,drop\_first=True, dtype=int)

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

# Among the numerical variables(temperature, humidity, windspeed), temperature has the highest correlation with the target variable(count).

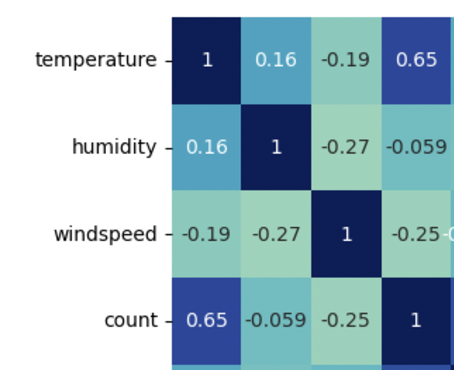
# The correlation coefficient of temperature and count is 0.65, which is highest among all other variables. Also, the regression line fits well for temperature as x and count as y.

**A comparison of the same graph

Description automatically generated with medium confidence**

**A graph showing the amount of temperature

Description automatically generated**

****

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of linear regression are validated using residual analysis, below are the assumption validations:

1. **Linearity**: The relationship between the independent and dependent variables is assumed to be linear.
2. Error terms are **normally distributed**: Plot a histogram/displot of the error terms and check whether the error terms are represented by normal distribution curve with mean as 0.

A graph of error terms

Description automatically generated

# Error terms are independent: There should be no visible pattern. Residuals (the differences between observed and predicted values) should be independent of each other.

A diagram of blue dots

Description automatically generated

1. **Homoscedasticity/** Error terms have **constant variance**: The variance of residuals should be constant, the variance should not increase/decrease as the error value changes.

A graph with blue dots

Description automatically generated

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

# Temperature, Year\_2019 and Month\_ September are the top three features contributing significantly towards explaining the demand of the shared bikes.

If we broadly classify the features by combining the overlapping features. Significant categories to predict the demand for shared bikes:

* Year: Year 2019 has higher bookings compared to year 2018, means the bookings are increasing year on year and company can predict higher booking for upcoming year.
* Days: Holiday time has least bookings whereas most working days and even weekends has a positive demand. Surprisingly, Mondays have less bookings which could be because of hybrid work culture where most companies are work from home on Mondays.
* Temperature/Weather Conditions or Season/Months:
  + The demand also increases with temperature (clear weather) and decreases for cloudy weather and highly decrease for bad weather condition with high windspeed, hence it should keep track of the weather conditions.
  + The month September and October (Fall season) should be considered by the company as they have a higher demand, this could be because of mostly clear weather condition. And Month January and February (Winter season) has negative impact on bookings which could be due to the severe weather conditions like snow and low temperature. Also July has less bookings which could be due to rainfall and cloudy weather condition

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 6 goes here>

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is used for predicting the value of the dependent variable based on the values of one or more independent variables. The basic idea is to find the best-fitting line (or hyperplane in

the case of multiple independent variables) that minimizes the sum of the squared

diﬀerences between the observed and predicted values of the dependent variable.

Linear regression algorithm follows following steps:

1. **Model Representation**:
2. **Simple Linear Regression**: In the case of a single independent variable, the model is

represented as:

𝑦 = 𝑏0 + 𝑏1⋅ 𝑥 + 𝜀

where:

- 𝑦 is the dependent variable,

- 𝑥 is the independent variable,

- 𝑏0 is the y-intercept (constant term),

- 𝑏1 is the slope of the line, and

- 𝜀 represents the error term.

1. **Multiple Linear Regression:** When there are multiple independent variables, the

model is extended to:

𝑦 = 𝑏0 + 𝑏1⋅ 𝑥1 + 𝑏2⋅ 𝑥2 + ⋯ + 𝑏n⋅ 𝑥n

where:

- (𝑥1, 𝑥2, … , 𝑥n) are the independent variables, and

- (𝑏0, 𝑏1, 𝑏2, … , 𝑏n) are the coeﬃcients.

1. **Objective Function:**

The goal is to find the values of (𝑏0, 𝑏1, 𝑏2, … , 𝑏n) that minimize the sum of the

squared diﬀerences between the observed and predicted values. This is

expressed as the sum of squared errors (SSE) or mean squared error (MSE).

1. **Minimisation:**

To find the optimal values of the coeﬃcients, the algorithm uses optimization

techniques such as gradient descent. The objective is to iteratively update the

coeﬃcients in the direction that minimizes the cost function.

1. **Training the Model:**

The model is trained on a dataset, where the algorithm learns the values of the

coeﬃcients that best fit the data. This involves feeding the algorithm input-output pairs and adjusting the coeﬃcients until the model produces predictions close to the

actual outcomes.

1. **Prediction:**

Once the model is trained, it can be used to make predictions on new data.

The predicted values are obtained by plugging the new input values into the learned regression equation.

1. **Evaluation:**

The model's performance is assessed using metrics such as (𝑅2) (coeﬃcient of

determination), MSE, or other relevant metrics, depending on the context.

1. **Assumptions:**

Linear regression relies on the assumption of a linear relationship between

independent and dependent variables, normally distributed errors, constant error

variance (homoscedasticity), and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

Linear regression is a versatile and widely used algorithm, but it's important to check whether its assumptions hold in each dataset and consider more advanced techniques when those assumptions are violated.

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but diﬀer significantly when graphed. This illustrates the importance of visualizing data and the limitations of relying solely on summary statistics. This quartet highlights the concept that datasets with similar statistical properties can exhibit diverse patterns when graphed.

A table of numbers and letters

Description automatically generated

All four sets are identical when examined using simple summary statistics but vary

considerably when graphed.

Mean\_x 9.000000 9.000000 9.000000 9.000000

Variance\_x 11.000000 11.000000 11.000000 11.000000

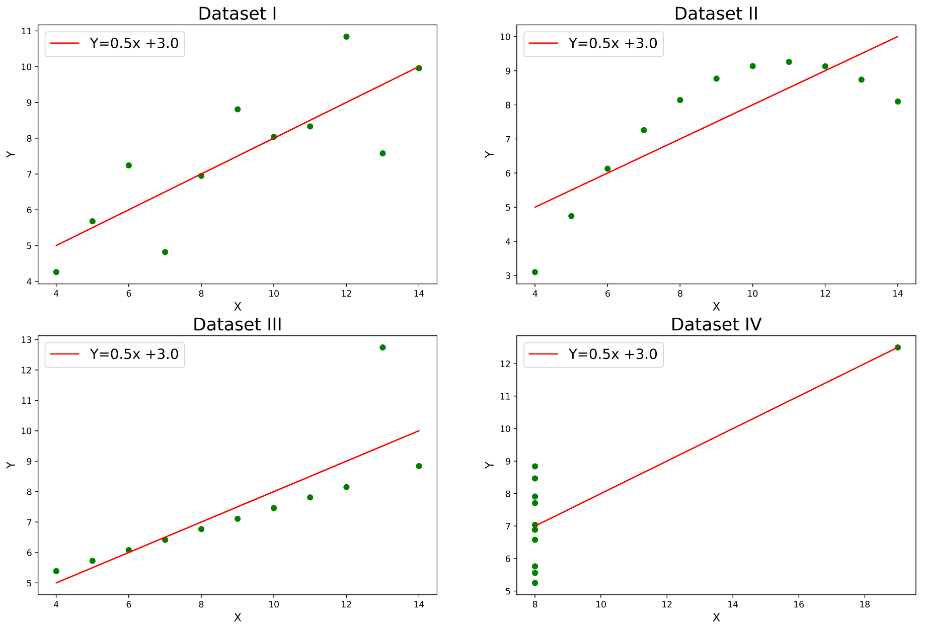
Mean\_y 7.500909 7.500909 7.500000 7.500909

Variance\_y 4.127269 4.127629 4.122620 4.123249

Correlation 0.816421 0.816237 0.816287 0.816521

Linear Regression slope 0.500091 0.500000 0.499727 0.499909

Linear Regression intercept 3.000091 3.000909 3.002455 3.001727



From the above diagram:

- The initial scatter plot (top left) suggests a straight forward linear relationship,

depicting two correlated variables, where y could be characterized as Gaussian with

a mean linearly dependent on x.

- In the second graph (top right), although a relationship between the variables is

evident, it is not linear, rendering the Pearson correlation coeﬃcient irrelevant. A

more general regression and the corresponding coeﬃcient of determination would

be more suitable.

- Moving to the third graph (bottom left), the modelled relationship is linear, but a

diﬀerent regression line is warranted (considering a robust regression). The

calculated regression is skewed by a single outlier, significantly reducing the

correlation coeﬃcient.

- Lastly, the fourth graph (bottom right) exemplifies a scenario where a lone high-

leverage point can yield a high correlation coeﬃcient, even when the other data

points fail to indicate any relationship between the variables.

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 8 goes here>

# The Pearson correlation coefficient (*r*) is a way of measuring linear correlation between two variables. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables, where:

- 𝑟 = 1: Perfect positive linear correlation.

- 𝑟 = − 1: Perfect negative linear correlation.

# - 𝑟 = 0: No linear correlation.

A value greater than 0 indicates a positive correlation i.e. as the value of one

variable changes, the value of other variable also changes in the same direction.

A value less than 0 indicates a negative correlation i.e. as the value of one variable changes, the value of other variable changes in the opposite direction.

The Pearson correlation coefficient also tells how close the observations are to a [line of best fit](https://www.scribbr.com/statistics/simple-linear-regression/#how-to-perform-a-simple-linear-regression) and whether the slope of the line of best fit is negative or positive. When *r* is 1 or –1, all the points fall exactly on the line of best fit and when *r* is 0, a line of best fit is not helpful in describing the relationship between the variables.

A graph of positive and negative

Description automatically generated

A diagram of a graph

Description automatically generated

## When to use the Pearson correlation coefficient:

* Both variables are [quantitative](https://www.scribbr.com/methodology/types-of-variables/#quantitative-vs-categorical)
* The variables are [normally distributed](https://www.scribbr.com/statistics/normal-distribution/)
* The data have no [outliers](https://www.scribbr.com/statistics/outliers/)
* The relationship is linear

The formula for Pearson's correlation coeﬃcient between two variables, 𝑋 and 𝑌, with 𝑛 data points, is given by:

The Pearson correlation for two objects, with paired attributes, sums the product of their differences from their object means, and divides the sum by the product of the squared differences from the object means.



It's important to note that correlation does not imply causa>on, and a correlation coeﬃcient close to zero does not necessarily mean the absence of a relationship; it only indicates the absence of a linear relationship.

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 9 goes here>

**Feature scaling** is one of the important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Scaling refers to the process of transforming the values of variables to a specific range or distribution. The goal is to bring all variables to a similar scale, making them comparable and preventing one variable from dominating others.

**Advantages of Scaling:**

1. **Equal Weightage**: Scaling ensures that all variables contribute equally to the

analysis, preventing variables with larger magnitudes from disproportionately

influencing the results.

2. **Convergence**: Many machine learning algorithms, particularly those based on

distances or gradients (e.g., k-nearest neighbours, support vector machines, gradient

descent-based algorithms), perform better when features are on a similar scale.

Scaling aids in faster convergence during the optimization process.

3. **Interpretability**: It improves the interpretability of coeﬃcients in linear models, as

the coeﬃcients represent the change in the dependent variable for a one-unit change in the predictor variable.

**Difference between normalized scaling and standardized scaling:**

**Normalization/Min-Max Scaling:**

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

A math equation with a number and a symbol

Description automatically generated with medium confidence

- Advantages: Useful when the distribution of the variable is unknown or not Gaussian.

- Disadvantages: Sensitive to outliers.

**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) as 0 and standard deviation (σ) as 1. sklearn.preprocessing.scale helps to implement standardization in python.

A black text on a white background

Description automatically generated

- Advantages: Less sensitive to outliers; preserves the shape of the distribution.

- Disadvantages: Assumes that the variable follows a Gaussian distribution.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a

multiple regression analysis. It quantifies how much the variance of the estimated

regression coeﬃcients is increased due to multicollinearity. The formula for VIF for a variable 𝑋i is:

VIF(𝑋i) = 1/(1-Ri ²)

Now, when you’re calculating the VIF for one independent variable using all the other independent variables, if the R² value comes out to be 1, the VIF will become infinite. This is quite possible when one of the independent variable is strongly correlated with many of the other independent variables.

When the value of VIF is infinite, it usually indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly correlated (linearly dependent) with other variables which is redundant information.

To address this issue, it's crucial to identify and handle multicollinearity in the

dataset. This can involve removing one of the perfectly correlated variables,

combining them or using dimensionality reduction techniques. Addressing

multicollinearity not only resolves the infinite VIF problem but also improves the

stability and interpretability of the regression model.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 11 goes here>

A quantile-quantile (Q-Q) plot is a graphical tool that compares two sets of data to determine if they come from the same distribution and determines if dataset follows a particular theoretical distribution, such as the normal distribution.

 Q-Q plots are useful in linear regression because they can help you check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals.

The Quantile-Quantile plot is used for the following purpose:

* 1. **Assessing Distributional Assumptions**: By comparing the quantiles of the observed data to the quantiles of the assumed distribution, deviations from the assumed distribution can be detected which impacts the accuracy of statistical inferences.
  2. **Detecting Outliers**: Q-Q plots can help identify outliers by revealing data points that fall far from the expected pattern of the distribution or deviate significantly from the rest of the dataset.
  3. **Comparing Distributions**: Q-Q plots can be used to compare two datasets to see if they come from the same distribution. This is achieved by plotting the quantiles of one dataset against the quantiles of another dataset. If the points fall approximately along a straight line, it suggests that the two datasets are drawn from the same distribution.
  4. **Assessing Normality**: Q-Q plots are particularly useful for assessing the normality of a dataset. If the data points in the plot closely follow a straight line, it indicates that the dataset is approximately normally distributed.
  5. **Model Validation**: Q-Q plots are used to validate predictive models. By comparing the quantiles of observed responses with the quantiles predicted by a model, one can assess how well the model fits the data.

**Interpretation of Q-Q Plots:**

- If the points in the Q-Q plot closely follow a straight line, it suggests that the

residuals are approximately normally distributed.

- Deviations from the straight line indicate departures from normality.

A graph with a line

Description automatically generated