
Assignment-based Subjective Questions

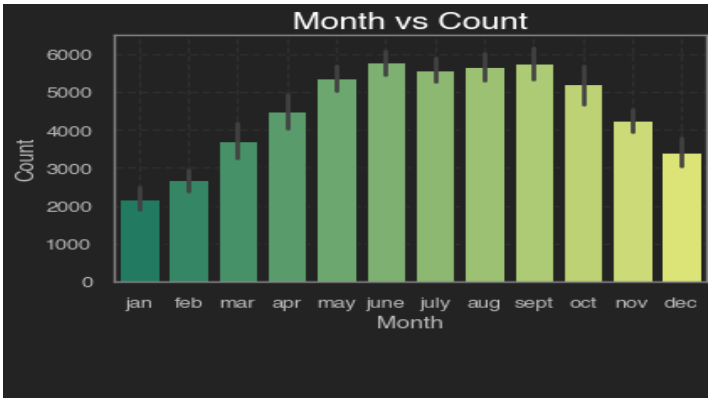
---

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

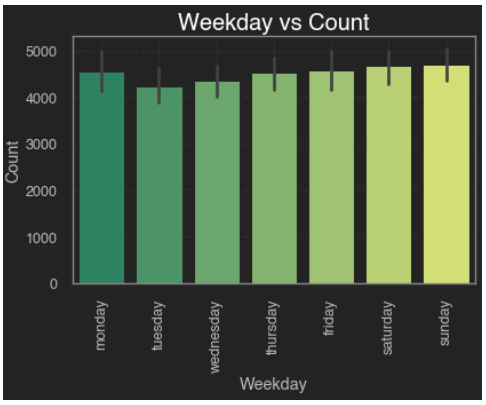
Ans:

- There are 5 Categorical Variables in the given Data set.Categorical Variables,
- X = Month, Weekday, Working Day, Season, Weathersit.Dependent Variable, y = Count.

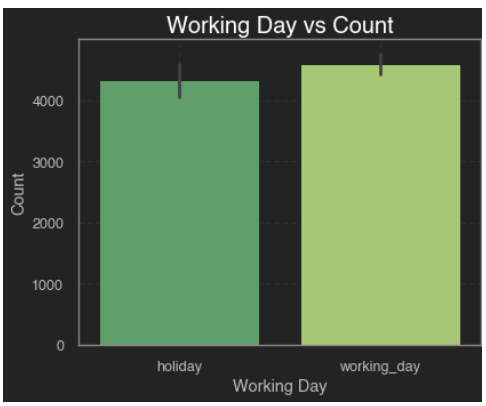
→ **Month:** June and September have the highest bike sharing Count.



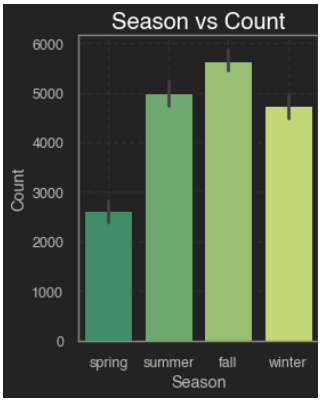
→ **Weekday:** Monday and Sunday have highest bike sharing Count.



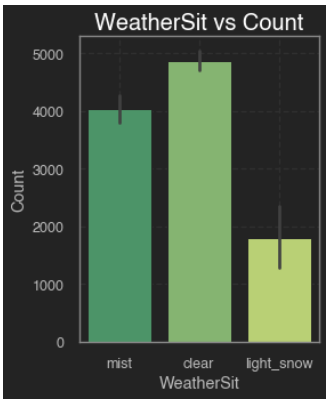
→ **Working Day:** Working Day has the highest Count.



→ **Season:** Fall has the highest bike sharing Count.



→ **Weather:** Clear weather has the highest bike sharing Count.



There is no demand for bikes in Heavy Rains.

---

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

Ans:

→ The **drop_first=True** will help to drop the last column that will be created using dummies method. This dropping of an extra column for each variable will decrease the correlations across the dummy variables.

→ While creating dummies for Categorical variable, the process will generate individual columns for each value for that variable.

→ Let's take an example using Season Categorical variable.

- Season Variable have 4 values:
  - Spring
  - Summer
  - Fall
  - Winter

→ So, dummies method will create 4 individual columns for each value.

- If we know three values from Season variable as,
  - Spring
  - Summer
  - Fall
- Then it is obvious that 4th value will be Winter.
- For all 4 values, it will create columns like below.

Spring	Summer	Fall	
1	0	0	1,0,0 is Spring
0	1	0	0,1,0 is Summer
0	0	1	0,0,1 is Fall
0	0	0	0,0,0 is Winter obviously

→ Here, we do not need all 4 columns:

- If we have n columns, then we need n-1 columns.
- Likewise, if we have 4 columns, we need 4-1 i.e., 3 columns.

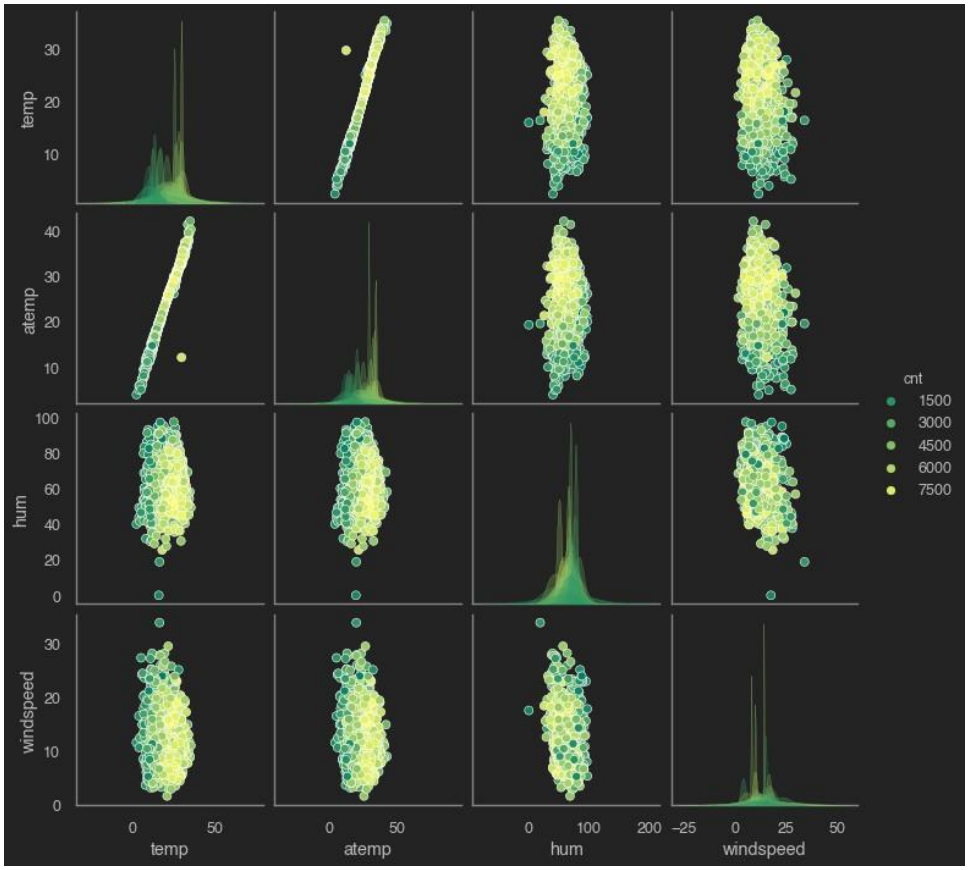
---

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

Ans:

→ Pair Plot:

- Pair Plot among numerical variables.
- With target variable **cnt** as hue.
- The variables **temp** and **atemp** are highly correlated with target variable.



→ Correlation Matrix:

- Heat Map among numerical variables including target variable.
- The variables **temp** and **atemp** are highly correlated with target variable.

---

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

Ans:

→ **Variance Inflation Factor (VIF):**
- All Variables have VIF less than 5.
- Hence, No Multicollinearity observed.

→ **Residual Analysis:**
- Residuals are normally distributed.

→ **Probability Plot of Residuals:**
- Residuals have linear relationship.

→ **Error Terms:**
- There is no visible pattern between error terms.
- Hence, Error Terms are Independent.

→ **Linearity Check:**
- Predicted values have Linearity with Actual values.

→ **Homoscedasticity:**
- The variance of residuals or error terms are Constant.
- The error terms not varying much with predicted values.

→ **No Multicollinearity:**
- There are almost no points that cross the boundary.
- Hence, No Multicollinearity for residuals.

→ **Actual vs Predicted:**
- There is minute difference between test and train data.

---

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

Ans:

→ **Temperatures:**
- Demand is high at Feel Like temperatures.

→ **Seasons:**
- Spring and Winter seasons have highest demand.

→ **Month:**
- January, February, September, November and December have highest demand.
- Hence, starting and ending of the year have highest demand.

General Subjective Questions

---

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

Ans:

→ Linear Regression is a Supervised Machine Learning Algorithm.
→ Regression is used to build model for predicting target variable based on independent variables.
→ This algorithm is used to predict values in a continuous range.
→ There are two types of Linear Regression Algorithms,
- Simple Linear Regression:
  - Dependent Variable which is linearly related to only One Independent variable.
  - It is represented by the equation,
    - $y = mx + c$
    - $Y = \beta_0 + \beta_1 x_1$
  - It is used to find best fit for coefficients, β_0 and β_1.
- Multiple Linear Regression:
  - Dependent Variable which is linearly related to more than One Independent variable.

→ Assumptions using Linear Regression,
- Variance Inflation Factor
- Residual Analysis
- Probability plot for Residuals
- Error Terms
- Linearity Check
- Homoscedasticity
- Checking Multicollinearity
- Actual and Predicted value differences

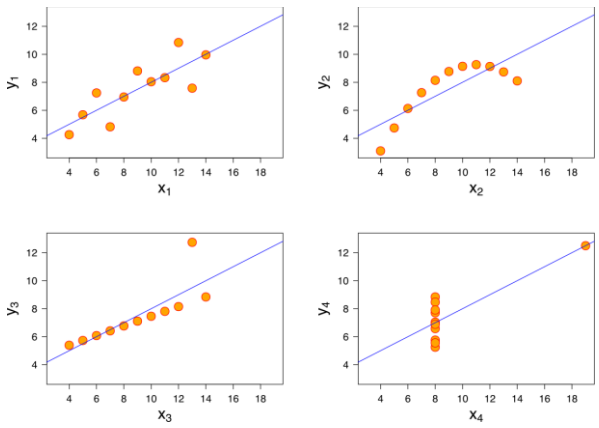
---

2. **Explain the Anscombe's quartet in detail.**

Ans:

Anscombe's quartet is a group of four data sets which provide useful care for applying individual statistical methods to the data without graphing them initially. They all have virtually the identical statistical properties, however they look totally different when they were plotted on a graph.

The data sets have 11 data points in each of the graphs which are identical in terms of their mean, correlation and variance. However, despite having same basic statistical properties, they look completely different on the plotted graph.



---

3. **What is Pearson's R?**

Ans:

→ Pearson's R is useful for measuring the strength of relationship of two variables which are on the same scale or interval.
→ It is always between –1 and 1. A value of 1 represents a perfect positive relationship which indicates that both variables move in the same direction.
→ On the converse, -1 represents perfect negative relationship which indicates that the variables move in the opposite directions, i.e., if one variable increases other decreases and vice-versa.
→ Zero value indicates no relationship between the two variables.

---

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

Ans:

→ Scaling is a step in the Data pre-processing which is generally applied to the independent variables to normalize the data within a given range.
→ It is also useful in speeding up the calculations while working on an algorithm.
→ Scaling is performed because data sets contain features of varying magnitudes, range and units. If scaling is not performed then the algorithm will only take magnitude into account and will not consider unit which may result in incorrect data modelling.
→ Hence, to resolve this issue, scaling is performed to bring all the variables in the data to the same level of range or magnitude.
→ Normalized Scaling is used to bring the all the variables in the range of zero and one whereas Standardized Scaling is used to scale all the values with mean (zero) and standard deviation (one).

---

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

Ans:

→ VIF = infinity when there is a high and direct correlation between two independent variables.
→ R2 = 1, when there a direct correlation,
  - 1/(1-R2) is infinity.
→ To overcome this situation, we need to drop the variables which are factors of this high multicollinearity.
→ VIF can also occur when one variable is directly related to the linear combination of remaining variables.

---

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

Ans:

→ Quantile is the percentage of points below a given value in data set.
→ For example, Median is a Quantile where 50% of data fall below that point and 50% lie above it.
→ The full form of Q-Q Plot is Quantile-Quantile plot.
→ Q-Q plot is scatter plot of two sets of quantiles against each other.
→ These plots are used to find out if two sets of data fetch from the same distribution.
→ Following problems can be addressed using Q-Q plot:
  - If 2 data sets fetched from same distribution.
  - If distributions having same shape.
  - If distributions having same skewness.
  - If two sets having same location and scale.

→ If two sets fetched from common distribution, a 45-degree angle will be generated.
→ Those points lie on the same line,
    - y = x