BIKE RENTAL REGRESSION ANALYSIS

MA 4142

Group 11

-Presentation By

- **Aabisheg T ES21BTECH11001**
- **Ansh vardhan ES21BTECH1108**
- **S Jagadeesh ES21BTECH11026**

Problem Statement:

- Rental bikes are introduced in urban cities to enhance mobility comfort.
- Timely availability and accessibility of rental bikes to the public are crucial to reduce waiting times.
- Ensuring a stable supply of rental bikes becomes a major concern for city management.
- The challenge is to effectively manage the distribution and availability of rental bikes.
- The goal is to meet public demand while optimizing operational efficiency and user satisfaction.

About Dataset:

```
> print(names(df))
 [1] "date"           "bike_count"      "hour"           "temp"
 [5] "humidity"       "wind_speed"      "visibility"     "dew_point_temp"
 [9] "solar_radiation" "rainfall"       "snowfall"       "seasons"
[13] "holiday"        "functioning_day"
```

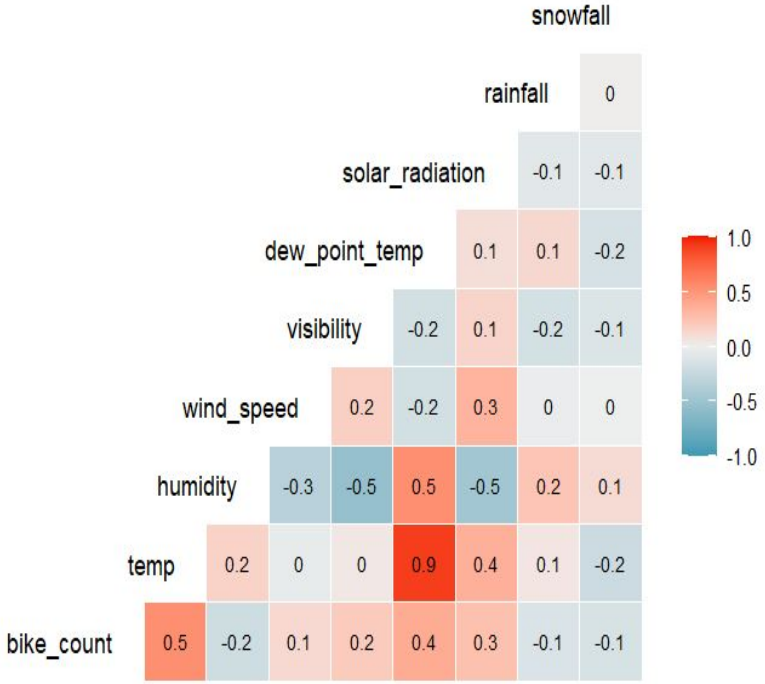
- **Numerical Variables:** Hour, Temperature, Humidity, Wind_Speed, Visibility, Dew_Point, Solar_Radiation, Rainfall, Snowfall.
- **Categorical Variables:** Season, IsHoliday, IsFunctioningDay.
- **Target Variable:** Bikes_Rented

Exploratory Data Analysis

In the exploratory data analysis (EDA) stage of data analysis, we look closely at the variables in the dataset in an effort to find any patterns or relationships that could point to correlations between them.

- **Correlation analysis.**
- **Categorical Variables analysis.**

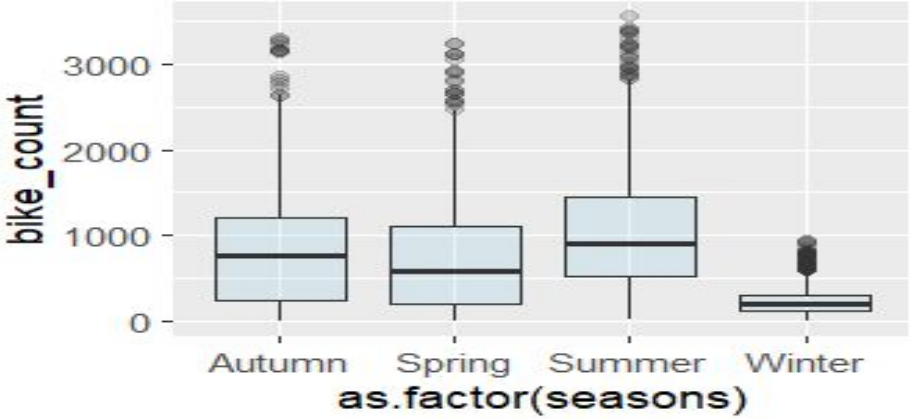
Correlation Analysis(for numerical data):

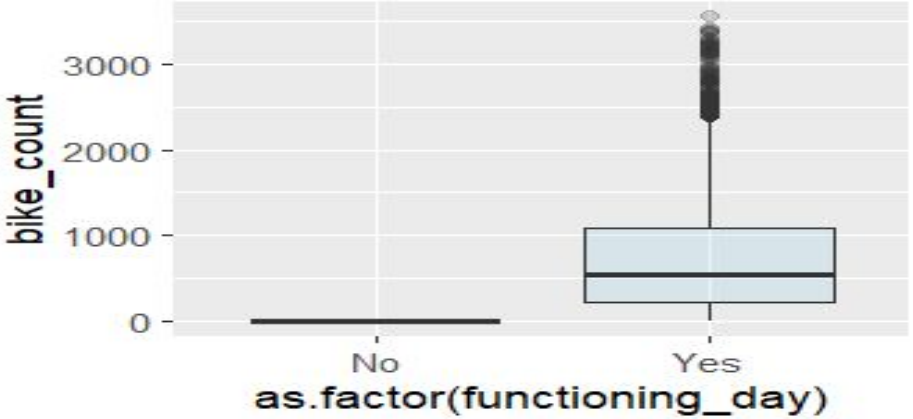


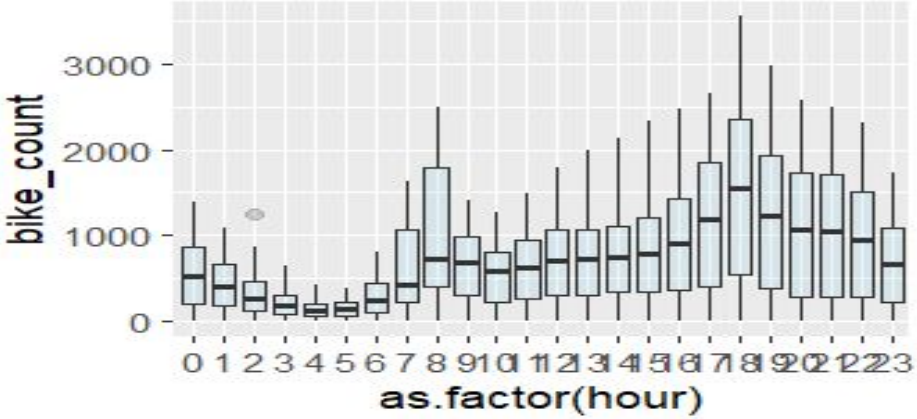
Based on the heatmap, the following pairs of variables exhibit high collinearity:

- temp (temperature) and dew_point_temp have a correlation of 0.9, which is a strong positive correlation.
- humidity and dew_point_temp have a correlation of 0.5, which is a moderate positive correlation.
- bike_count and temp show a correlation of 0.5, which is a moderate positive correlation.

Categorical Variables Analysis:







Observations :

- Most of the bike reservations were made on functioning days indicating a significant bias in the statistics.
- The boxplot shows that the bike count is always 0 on non-functioning days since bikes are not available for hire on those days.
- As a result, since they are unrelated to our research, data points for days when the system was not operational are deleted from the dataset.Following this filtering, there are 8465 data points in the dataset.
- Also the box plot shows that more bikes were rented on a working day(non-holiday).

Modelling:

Train-test split: Prior to constructing the model, it's essential to divide the dataset into two separate subsets: the training dataset and the test dataset. The plan is to allocate 70% of the data to the training dataset, while the remaining 30% will be designated as the testing dataset.

Regression(Model 1): In this model we will try to model the linear regression using 'the number of bikes rented per hour' as the target variable and all other variable as the predictors.

```
model_1 <- lm(bike_count ~ ., data = data_train)
```

Summary of Model 1:

Residual Standard Error: 375 Indicates average deviation of observed values from fitted values

R^2 : Multiple - 0.6666, Adjusted - 0.6647

Explains approximately 66.66% of variability in dependent variable

Adjusted for model complexity

F-Statistic: 336.5

Tests overall significance of the model

Extremely small p-value (< 2.2e-16) indicates strong evidence against null hypothesis

```
Residual standard error: 375 on 5890 degrees of freedom
Multiple R-squared:  0.6666,    Adjusted R-squared:  0.6647
F-statistic: 336.5 on 35 and 5890 DF,  p-value: < 2.2e-16
```

Removing insignificant features from the model using step regression analysis:

```
stepReg = MASS::stepAIC(model_1 , direction = 'both')
stepReg$anova
```

Stepwise method produced a final regression model without visibility, and wind speed. These will be removed from the final model.

```
Initial Model:
bike_count ~ hour + temp + humidity + wind_speed + visibility +
    dew_point_temp + solar_radiation + rainfall + snowfall +
    seasons + holiday

Final Model:
bike_count ~ hour + temp + humidity + dew_point_temp + solar_radiation +
    rainfall + snowfall + seasons + holiday
```

Model with significant features:

```
bike_count ~ hour + temp + humidity + dew_point_temp + solar_radiation +
    rainfall + snowfall + seasons + holiday
```

Summary of this model:

```
Residual standard error: 374.9 on 5892 degrees of freedom
Multiple R-squared:  0.6666,    Adjusted R-squared:  0.6647
F-statistic:   357 on 33 and 5892 DF,  p-value: < 2.2e-16
```

With an adjusted R-squared of 0.6647, the original model including all predictors can account for 66.47% of the target variables. but the corrected R-squared for the second model is also 0.6647. this implies that a model with just significant predictors can be used.

Model's Performance:

```
    RMSE(pred = model_1_sig$fitted.values, obs = data_train_sig$bike_count)
373.8531
    RMSE(pred = pred_sig, obs = data_test_sig$bike_count)
374.1233
```

Performing Tests on Model:

Shapiro Wilk's Test for Normality:

```
        Shapiro-Wilk normality test

data:  model_1_sig$residuals[0:5000]
W = 0.98536, p-value < 2.2e-16
```

Since P-value < 0.05 so H0 is rejected or the residual does not follow a normal distribution

BP Test on Homoscedasticity:

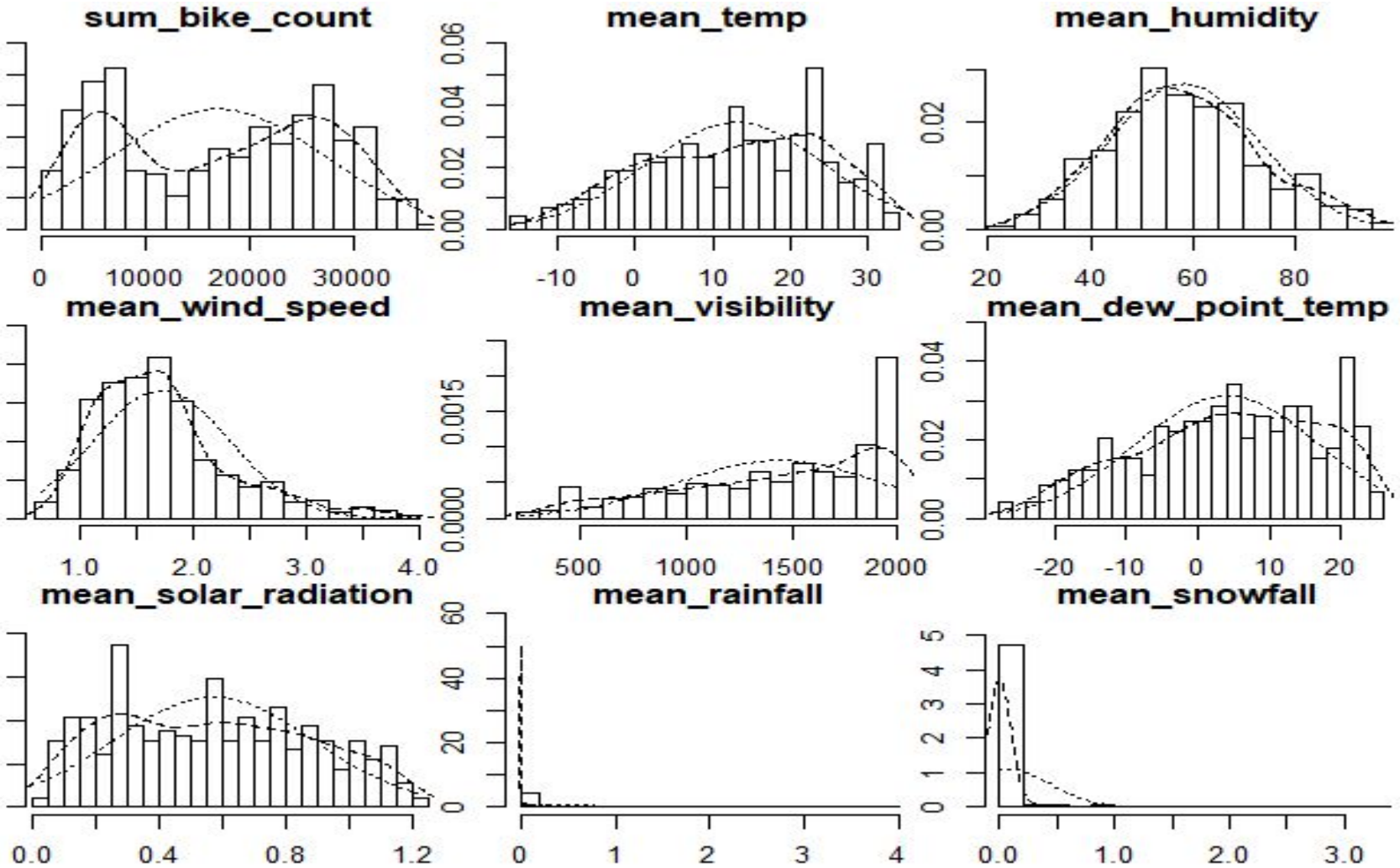
```
        studentized Breusch-Pagan test

data:  model_1_sig
BP = 1721.7, df = 33, p-value < 2.2e-16
```

With a p-value <0.05, we can conclude that there is heteroscedasticity in our model.

We observe that our model fails both Normality Test and Homoscedasticity, so still need to move to a better model.

New_model (grouped by dates):

- In this model we have grouped data based of dates(i.e for a given date, average values of data for that date are taken).
- The histograms showing the continuous variables' appearance following data transformation are shown below.
- In comparison to the original data, the temperature shows more of a flatter distribution while the sum_bike_count indicates signs of having a bimodal distribution. The remaining columns appear to be somewhat similar to the first distribution of data..


sum_bike_count
0
10000
20000
30000
0.00
0.02
0.04
0.06
mean_temp
-10
0
10
20
30
mean_humidity
20
40
60
80
mean_wind_speed
1.0
2.0
3.0
4.0
0.0000
0.0015
mean_visibility
500
1000
1500
2000
mean_dew_point_temp
-20
-10
0
10
20
0.04
mean_solar_radiation
0.0
0.4
0.8
1.2
0
20
40
60
mean_rainfall
0
1
2
3
4
mean_snowfall
0
1
2
3
4
5
0.0
1.0
2.0
3.0

Building new_model:

```
model_2 <- lm(sum_bike_count ~ mean_temp + mean_humidity +
mean_wind_speed + mean_visibility + mean_dew_point_temp +
mean_solar_radiation + mean_rainfall + mean_snowfall + seasons + holiday,
data=df_train)
```

Summary of this model:

```
Residual standard error: 3940 on 269 degrees of freedom
Multiple R-squared:  0.8501,    Adjusted R-squared:  0.8434
F-statistic: 127.1 on 12 and 269 DF,  p-value: < 2.2e-16
```

- With a p-value of less than 0.05, this model is statistically significant and able to explain 0.8378. It does, however, demonstrate that certain elements are not statistically significant. Once more, stepwise AIC is employed to choose the best model for prediction.

Removing insignificant features from new_model:

Performing Step Regression:

```
model_2_step = step(model_2 , direction = 'back')

summary(model_2_step)

Residual standard error: 3932 on 271 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8441
F-statistic: 153.2 on 10 and 271 DF,  p-value: < 2.2e-16
```

The stepwise regression results show a slight increase in adjusted R-Squared to 0.8382 by reducing predictors mean_visibility and mean_snowfall.

Model testing:

Normality test(Shapiro-Wilk's test):

```
shapiro.test(model_2_final$residuals[0:5000])

        Shapiro-Wilk normality test

data:  model_2_final$residuals[0:5000]
W = 0.99635, p-value = 0.7624
```

P-value of the Shapiro-Wilk is greater than 0.05 therefore, it can be concluded that normality error assumption is not violated.

Heteroscedasticity test(BP test):

```
bptest(model_2_final)

        studentized Breusch-Pagan test

data:  model_2_final
BP = 60.772, df = 9, p-value = 9.514e-10
```

p-value = 9.514e-10 < 0.05 so we have to reject H0. Residuals are not distributed evenly (heteroscedasticity)

Multicollinearity test:

```
vif(model_2_final)
                      GVIF Df GVIF^(1/(2*Df))
mean_temp         6.454181  1        2.540508
mean_humidity     2.223462  1        1.491128
mean_wind_speed   1.202904  1        1.096770
mean_solar_radiation 3.244988  1        1.801385
mean_rainfall     1.500984  1        1.225147
seasons           6.040279  3        1.349510
holiday           1.033729  1        1.016725
```

- Since all the GVIF values are less than 10, it can be assumed that there is no multicollinearity for the final Model.
- This model passes the test for normality and multicollinearity but it fails homoscedasticity test so we perform log transformation on response variable.

Log transformation:

```
df_average_log$sum_bike_count <- log(df_average$sum_bike_count + constant)
```

New_model after log transformation:

```
model_final_log <- lm(sum_bike_count ~ mean_temp + mean_wind_speed +
                        mean_solar_radiation + mean_rainfall + seasons + holiday,
                      data = df_train_log)

summary(model_final_log)
```

```
Residual standard error: 0.2825 on 238 degrees of freedom
Multiple R-squared:  0.8571,    Adjusted R-squared:  0.8523
F-statistic: 178.4 on 8 and 238 DF,  p-value: < 2.2e-16
```

After log transformation , clearly the value of Adjusted R-square increased to 0.8523 from 0.8441.

Performing all 3 tests on this model:

Normality test:

```
        Shapiro-Wilk normality test

data:  model_2_final$residuals[0:5000]
W = 0.99635, p-value = 0.7624
```

Heteroscedasticity test(BP test):

```
        studentized Breusch-Pagan test

data:  model_final_log
BP = 49.524, df = 8, p-value = 5.045e-08
```

As we can observe we still didn't pass the BP test but our p-value considerably improved from 9.514e-10 to 5.045e-08 which shows that Log Transformation helped our model.

Multicollinearity test:

	GVIF	Df	GVIF^(1/(2*Df))
mean_temp	5.388442	1	2.321302
mean_wind_speed	1.161436	1	1.077700
mean_solar_radiation	2.282712	1	1.510865
mean_rainfall	1.420053	1	1.191660
seasons	5.441546	3	1.326235
holiday	1.034016	1	1.016866

Conclusion

- The final model includes temperature, wind speed, solar radiation, mean rainfall, seasons, and holidays as predictors of bicycle rentals.
- The model has been assessed for several assumptions, including normality and multicollinearity, which it satisfies. However, it fails the Homoscedasticity of Residuals assumption.
- Despite the violation of the Homoscedasticity assumption, the model demonstrates a high R-squared value, indicating that approximately 85% of the variability in bicycle rentals can be explained by the included variables.
- The model's predictive accuracy is evaluated using the root mean squared error (RMSE). The RMSE for the training dataset is 48.83893, and for the testing dataset, it is 14.87956.

Thank You!