**Introduction to Business Decision Process**

**DSCI 5180**

**Star Dataset**

Student Name: Donetra Hemraj Lanjewar

Student ID: 11561188

Student mail Id: donetrahemrajlanjewar@my.unt.edu

The illuminating celestial bodies are always fascinating. The data revolves around Stars and its types. It covers the related aspects that can be used to determine the type.

Variables that are included in the dataset involves:

* Temperature ( in Kelvin)
* Luminosity
* Absolute Magnitude
* Radius
* Spectral class
* Star colour
* Star type

Spectral class, Star colour, star type are further divided into sub-categories as stated below:

* Spectral Class – M, A, B, F, K, O, G
* Star Colour – Red, White, Blue – white, Blue, Yellowish – white, Pale – yellow – orange, White – yellow, Whitish, Yellow – white, Yellowish, Orange – red, Orange
* Star type – 0, 1, 2, 3, 4, 5

The dataset contains 240 entries of data with 40 for each star type.

Source : Kaggle

Datalink : <https://www.kaggle.com/datasets/deepu1109/star-dataset>

The project starts with descriptive statistics and ends with predictive. Modules covered in this include :

* Probability calculation in Standard Normal Distribution
* Estimating the Confidence Interval for Population Mean
* Hypothesis Testing with unknown Standard Deviation
* Multiple Regression Model
* Inferences concerning the multiple regression model and it’s coefficients.

The starts always appear calm and soothing to our eye. But spitting out the facts, it’s very hot with temperature being ranged more than thousands of Kelvin. The light emitted by stars is also affected by its temperature.

Temperature being a very common parameter with most importance in Stellar, in the last section of this project, temperature is predicted using the regression model. **Module – 1**

The absolute magnitude of stars are normally distributed. The mean of absolute magnitude is 4.38 with population standard deviation being 10.51. Find the probability that the absolute magnitude of a randomly selected star is greater than 10.

Solution:

μ = 4.38 , σ = 10.51

P( X > 10) = P(z > 0.5347) = 0.703492862. ----------------*Please refer Appendix A.1*

**Module – 2**

A scientist desires to know the mean temperature of each star type. The sample of 40 each is drawn with the details as follows:

|  |  |  |
| --- | --- | --- |
| Star Type | Mean Temperature ( in Kelvin ) | Population Standard Deviation |
| 0 | 2997.95 | 328.10 |
| 1 | 3283.825 | 266.25 |
| 2 | 13931.45 | 4895.29 |
| 3 | 16018 | 10527.13 |
| 4 | 15347.85 | 9959.90 |
| 5 | 11405.7 | 116698.34 |

Construct 95% confidence interval for the mean temperature of stars among all type individually with respected information as stated above.

Solution:

Confidence interval for population mean is given by ‾x – E < μ < ‾x + E

Margin of error, E = z) \*

c = 0.95,α = 100-0.95 = 0.05

Critical value, z0.025 = 0.4920 ---------------------*Please refer Appendix A.2*

For type 0, E = 0101.6764454 ⇒ Confidence interval = (2896.27, 3099.63)

For type 1, E = 82.50880639 ⇒ Confidence interval = (3201.32,3366.33)

For type 2, E =1517.038913. ⇒ Confidence interval = (12414.41,15448.49)

For type 3, E = 3262.330215. ⇒ Confidence interval = (12755.67,19280.33

For type 4, E = 3086.548824. ⇒ Confidence interval = (12261.30,18434.40)

For type 5, E = 3615.990997. ⇒ Confidence interval = (7789.71, 15021.70 )

**Module - 3**

A claim is made that the stars have an average temperature less than 10000 K. To test this, a random sample is being drawn of 144 stars with mean of 12529.28 K, sample standard deviation is known to be 10854.89. The claim is tested with level 0.05 of significance.

Solution:

* Hypothesis Statement:

Null hypothesis : The mean temperature is 10000

Alternative hypothesis : The mean temperature is less than 10000

* The parameter used in this hypothesis test is mean temperature so the statistical measure is mean.
* With statistical measure, the hypothesis statement is stated as :

H0 : μ = 10000

Ha : μ < 10000

* From the given statement of test, the claim is stating that the mean is less than 10000.
* Therefore, it’s an one tailed test (Left tailed test).
* The level of accuracy as given is 95%

Therefore, α = 0.05

* Sample standard deviation is given. This implies it’s a t test. …..*Please refer appendix A.3*

Degree of freedom – Number of sample – 1 = 144 – 1 , df = 143

t critical = 1.65557914; p = 0.00294202; t statistical = -2.79

* If the *p-*value is less than or equal to α, we should reject the null hypothesis. If the *p*-value is greater than α, we fail to reject it. We know that α=0.05 and the *p*-value is 0.0029.

Here, p-value < 0.05, the test statistic falls in the rejection region.

Therefore, we reject null hypothesis.

There is sufficient evidence to conclude that the mean of temperature for stars is less than 10000.

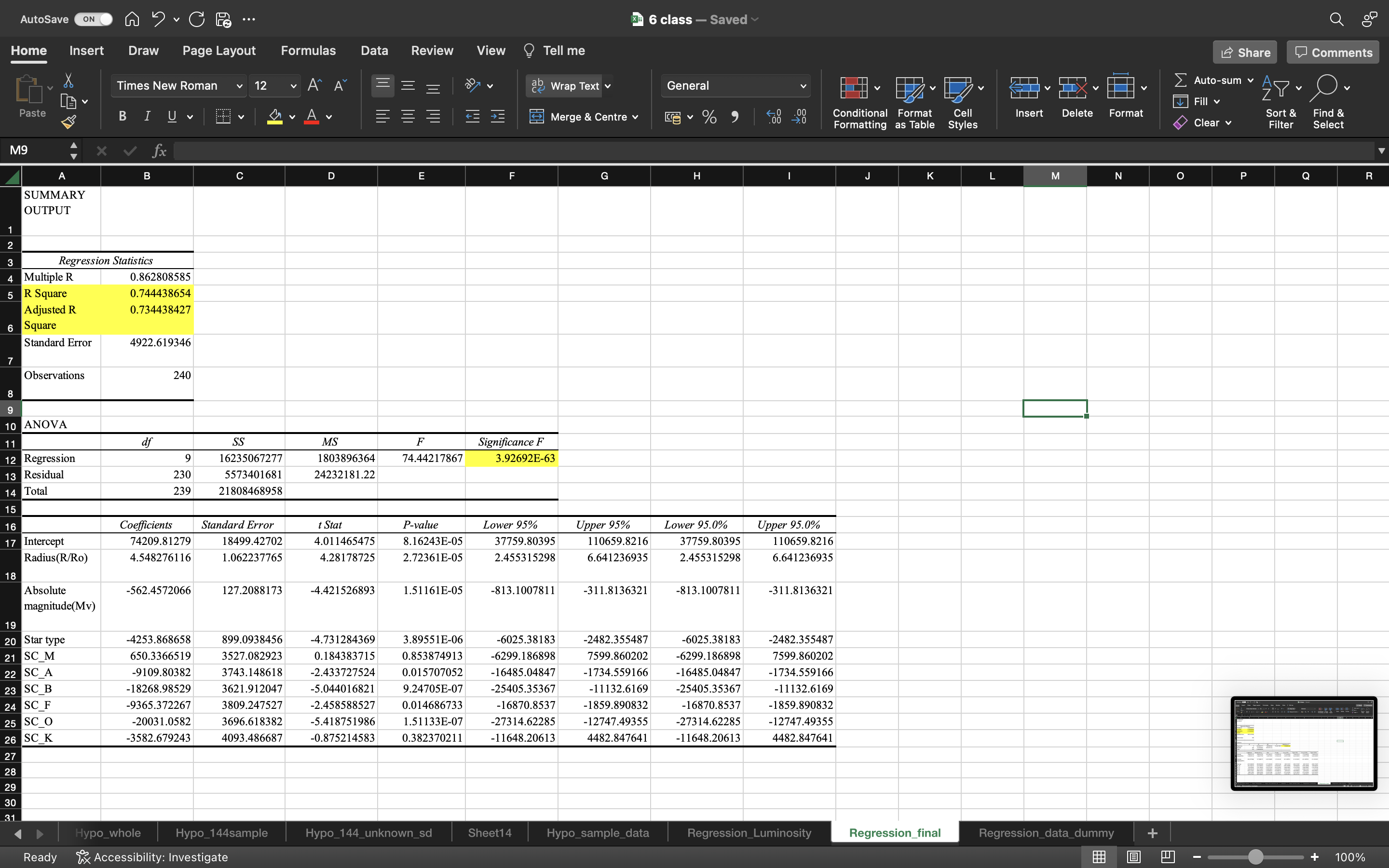
**Module – 5**

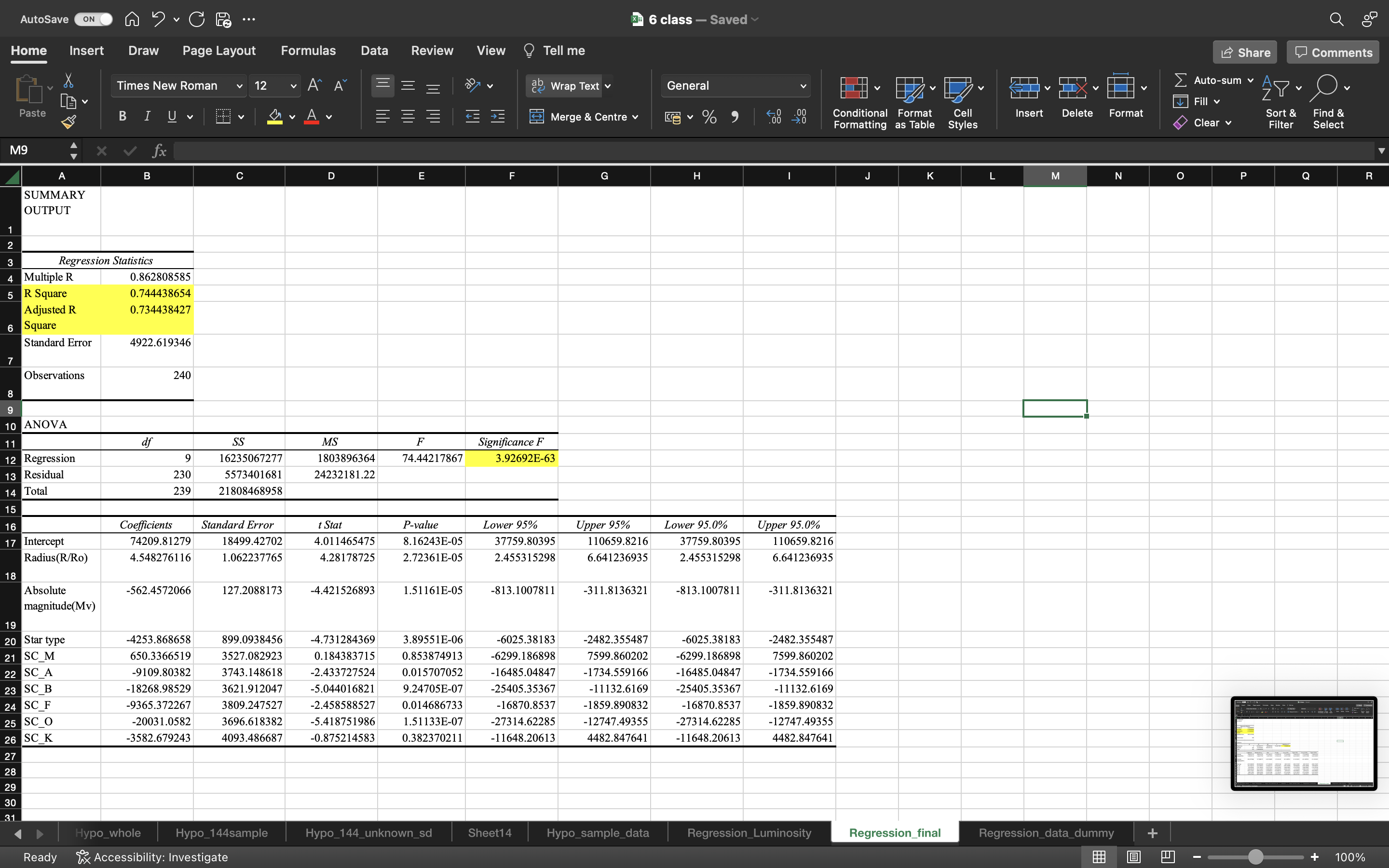
Regression model is used to predict the temperature for stars with the use of radius, absolute magnitude, star type and spectral class.

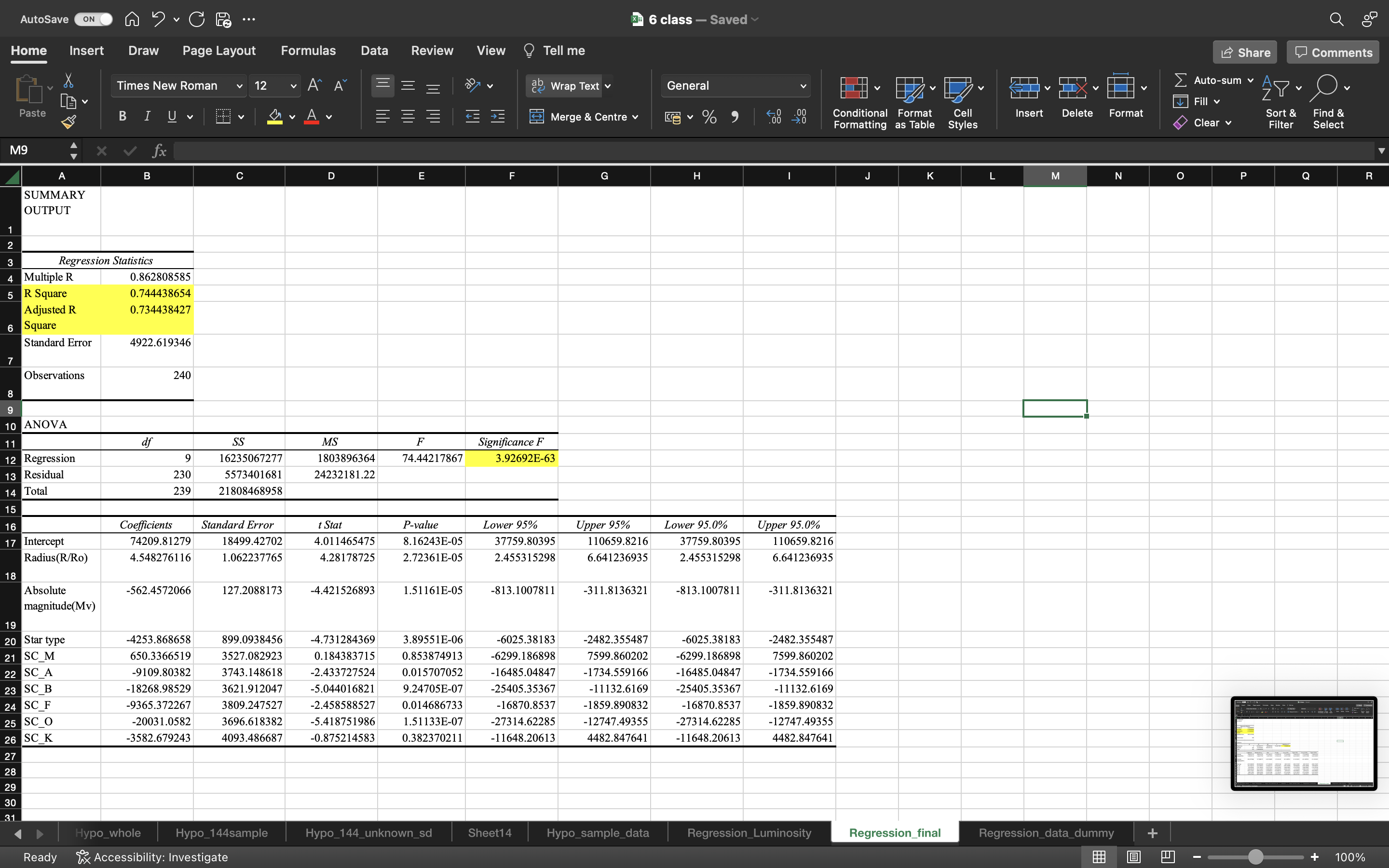
Solution:

The dataset is assumed to be normally distributed.

Also, dependent variable i.e., Temperature is a continuous variable

 …………………….*Please refer appendix A.4*





Temperature = 74209.81 + 4.55(Radius) - 562.46(Absolute Magnitude) - 4253.87(Star type) + 650.34(SC\_M) – 9109.8(SC\_A) – 18268.99(SC\_B) – 9365.37(SC\_F) – 20031.06(SC\_O) – 3582.68(SC\_K)

The regression equation is of the format

Y = bo + b1x1 + b2xx + b3x3 + b4x4 + b5x5 + b6x6 + b7x7 + b8x8 + b9x9

Where, bo = intercept

b1 = coefficient for Radius

b2 = coefficient for Absolute Magnitude

b3 = coefficient for Star Type

b4 = coefficient for SC\_M (dummy Variable for Spectral class)

b5 = coefficient for SC\_A (dummy Variable for Spectral class)

b6 = coefficient for SC\_B (dummy Variable for Spectral class)

b7 = coefficient for SC\_F (dummy Variable for Spectral class)

b8 = coefficient for SC\_O (dummy Variable for Spectral class)

b9 = coefficient for SC\_K (dummy Variable for Spectral class)

Spectral class being categorical variable, is converted into dummy variable which can be used in the regression model

Variation in temperature as expressed by the regression is 16235067277.

Proportion of variation explained by regression = 0.744.

Estimated variance of the error terms = 24232181.22.

Total variability of the dependent terms = 21808468958.

Variance of dependent variable = 91248824.09.

Percent of variation in estimated variable as explained by the independent variable = 74.43 %

**APPENDIX**

**A.1 Finding probability for X > 10**

μ = 4.38 , σ = 10.51 Given information

P( X > 10) lies in the shaded region below Problem statement

P( Converted to Standard Normal

P(z > 0.5347) = 1 - P(z < - 0.5347) = 1 – 0.2946

Since the normal curve is symmetric

P(z > 0.5347) = 0.703492862

**A.2 Confidence Intervals**

Confidence interval for population mean is given by ‾x – E < μ < ‾x + E

Where, ‾x is sample mean and E is margin of error

E = z) \*

We get confidence interval as (‾x + E , ‾x - E )

Where , Upper limit =‾x + E, and

Lower limit =‾x - E

Given that level of confidence is 95%, c = 0.95, we can determine that α = 100-0.95 = 0.05

Therefore we can get the critical value from the table for z0.025 = 0.4920

For type 0, E = 0.4920\* = 101.6764454

Lower limit = 2997.95 – 101.676

Upper limit = 2997.95 + 101.676

⇒ Confidence interval = (2896.27, 3099.63)

For type 1, E = 0.4920\* = 82.50880639

Lower limit = 3283.825 – 82.509

Upper limit = 3283.825 + 101.609

⇒ Confidence interval = (3201.32,3366.33)

For type 2, E = 0.4920\* = 1517.038913

Lower limit = 13931.45 – 1517.039

Upper limit = 13931.45 + 1517.039

⇒ Confidence interval = (12414.41,15448.49)

For type 3, E = 0.4920\* = 3262.330215

Lower limit = 16018 – 3262.33

Upper limit = 16018 + 3262.33

⇒ Confidence interval = (12755.67,19280.33

For type 4, E = 0.4920\* = 3086.548824

Lower limit = 15347.85 – 3086.549

Upper limit = 15347.85 + 3086.549

⇒ Confidence interval = (12261.30,18434.40)

For type 5, E = 0.4920\* = 3615.990997

Lower limit = 11405.7 – 3615.99

Upper limit = 11405.7 + 3615.99

⇒ Confidence interval = (7789.71, 15021.70 )

**A.3 Hypothesis Testing for μ < 10000**

H0 : μ = 10000

Ha : μ < 10000

The level of accuracy as given is 95%

Therefore, α = 0.05

Degree of freedom – Number of sample – 1 = 144 – 1

df = 143

From t – table,

t critical = 1.65557914;

p value for t critical is 0.00294202

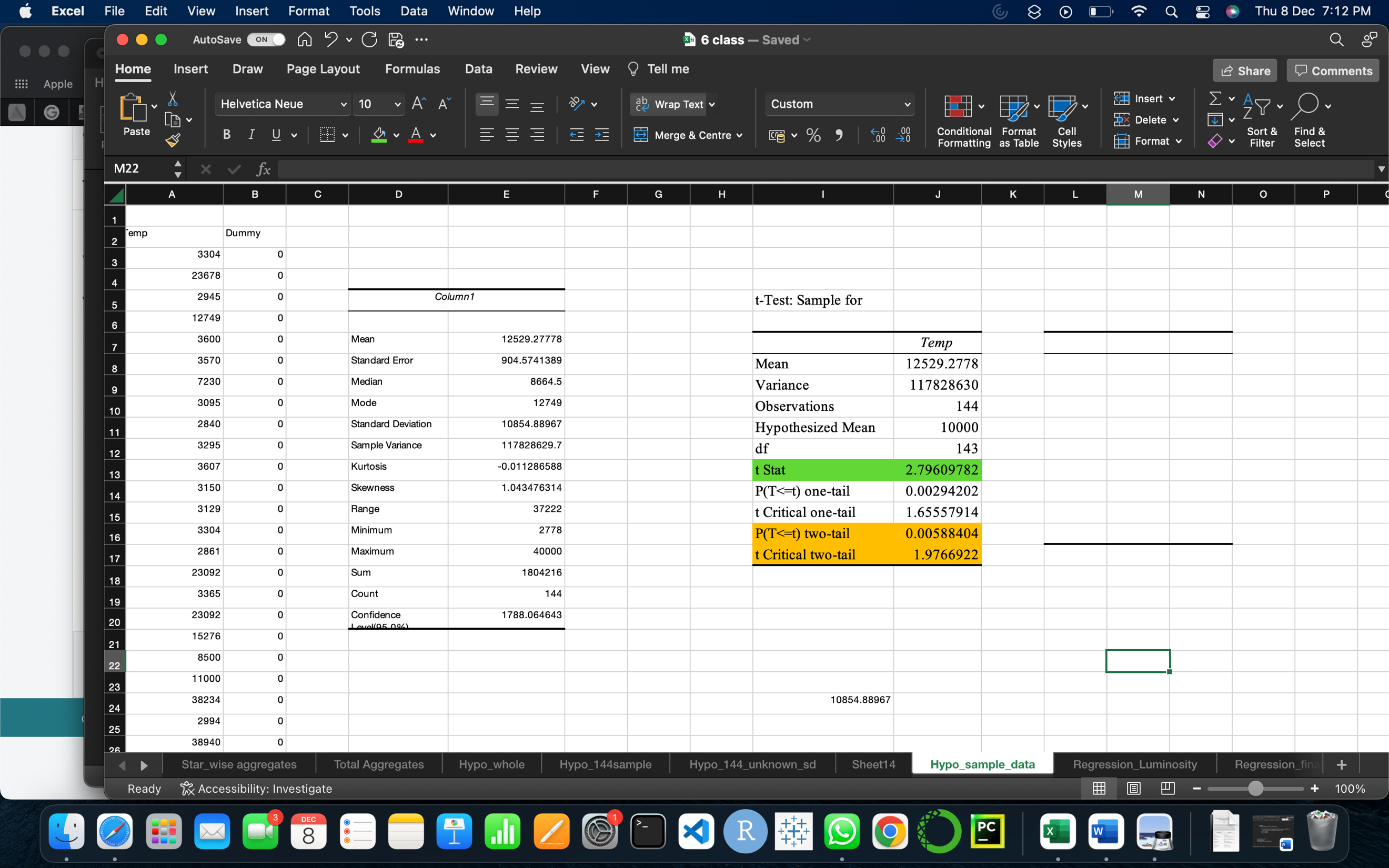
t statistical = = -2.79 …………………..*Please refer appendix A.3*

* If the *p-*value is less than or equal to α, we should reject the null hypothesis. If the *p*-value is greater than α, we fail to reject it. We know that α=0.05 and the *p*-value is 0.0029.

Here, p-value < 0.05, the test statistic falls in the rejection region.

Therefore, we reject null hypothesis.

There is sufficient evidence to conclude that the mean of temperature for stars is less than 10000.



**A.4 Regression Model**

Independent Variables : Radius, Absolute Magnitude, Spectral Class

Dependent Variable : Temperature

5 Dummy variables are created for spectral class (spectral class variables = 6) namely SC\_M, SC\_A, SC\_B, SC\_F, SC\_O and SC\_K.

The unique values from Spectral class are as follows :

|  |
| --- |
| Unique Values |
| M |
| A |
| B |
| F |
| K |
| O |
| G |

Here, n = 7

Based on this, ( n – 1), 6 dummy variables were created to proceed with the regression model.

Reference for this can be found in Sheet named Regression\_data\_dummy

Luminosity being discovered as useless parameter for temperature prediction, it was discarded. Reference for this can be found in sheet named Regression\_with\_luminosity.

Variation in temperature as expressed by the regression = SSR = 16235067277.

Proportion of variation explained by regression = = 0.744.

Estimated variance of the error terms = Mean square error (MSE) = 24232181.22.

Total variability of the dependent terms = TSS = 21808468958.

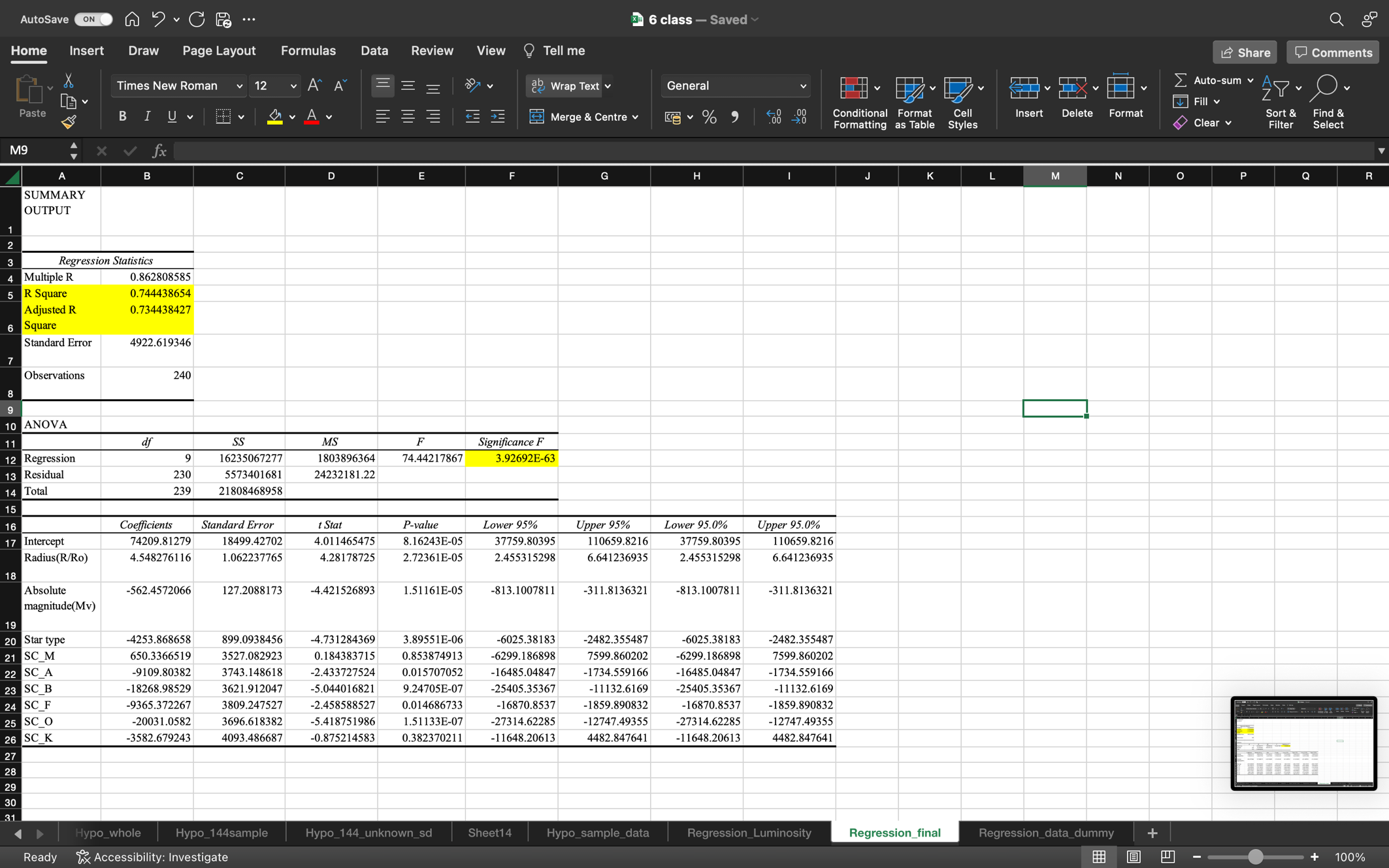
Variance of dependent variable = = 91248824.09.

Percent of variation in estimated variable as explained by the independent variable =

= = 74.43 %

The interpretation of ANOVA Table is as follows :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum of Squares (SS) | Degrees of freedom (df) | Mean Square (MS) | F - significance |
| Between  (Treatment) | SST | k – 1 |  |  |
| Within  (Error) | SSE | n – k |  |  |
|  | TSS | n – 1 |  |  |



The multiple regression equation :

Temperature = 74209.81 + 4.55(Radius) - 562.46(Absolute Magnitude) - 4253.87(Star type) + 650.34(SC\_M) – 9109.8(SC\_A) – 18268.99(SC\_B) – 9365.37(SC\_F) – 20031.06(SC\_O) – 3582.68(SC\_K)

A.4s References in Excel Workbook

For Question 1 – Sheet 1 (Total Aggregates)

For Question 2 – Sheet 2 (Star\_wise aggregates

For Question 3 – Sheet 3 (Hypo\_sample\_data)

For Question 4 – Sheet 4 and 5 (Regression\_final and Regression\_data\_dummy)