ALY 6015 Introduction to Enterprise Analytics

Submitted by: Bhavesh Patidar and Tanishq Bakliwal

Submitted to: Prof. Zhi He

Date: 10/20/2023

Introduction:

In our research project, we are delving into the fascinating world of meteorite landings, armed with a comprehensive dataset curated by Javier de la Torre and generously provided by The Meteoritical Society. This extensive dataset encompasses a wealth of information on 34,513 meteorites, allowing us to explore their various facets. From the meteorite type to its weight, and from whether it fell from the sky or was discovered, to the year of its landing, we have an array of key attributes at our disposal. The dataset also includes database information, and most notably, the latitude and longitude coordinates that pinpoint the exact locations of these celestial visitors' landings. Additionally, we have access to a trove of metadata fields, offering deeper insights into the meteorites' characteristics and origins. With this rich dataset as our foundation, we are poised to embark on an in-depth investigation into the mysteries of meteorite landings and their significance.

The goal of this project is to verify and analyze meteorite distribution, type and classification. Along with this we will find the impact of meteorite over the time and what proportion of meteorites were observed falling versus those discovered later.

In this project we have used three methods which will answer all the questions about meteorite landing. First method which we have used is logistics regression model. Logistic regression is a statistical model used for binary classification, which means it's used to predict one of two possible outcomes. It's called "logistic" because it uses the logistic function to model the probability of an event occurring.

Second method which we have used is multi variable method. A multivariable method, often called multivariate analysis, involves analyzing and understanding the relationships between multiple variables or factors simultaneously. Instead of looking at one variable in isolation, it considers how several variables interact and influence each other to gain a more comprehensive understanding of a situation or problem.

Third method which we used is linear regression model. A statistical method for quantifying the linear connection between a dependent variable and one or more independent variables is called a linear regression model. It gives an equation in the shape of a straight line that best matches the observed data and estimates the effect of changes in the independent variables on the dependent variable. In disciplines like data science, social sciences, and economics, this model is frequently used for predictive analysis, comprehending the correlations between variables, and hypothesis testing.

Methods:

The dataset which we are using gives us information about meteorite landings, including details like location, meteorite type, mass, year, and geographical coordinates. All three logistic regression, multivariable method and linear regression model can be useful for this dataset, but they serve different purposes:

Logistic Regression:
Linear regression is commonly used for predicting a continuous numeric target variable based on one or more input features. In the context of meteorite data, we could use linear regression to make predictions about a specific numeric attribute, such as predicting the mass of a meteorite based on other factors.

Multivariable Analysis:
Multivariable analysis is not a specific algorithm but a broad approach that involves studying relationships and patterns between multiple variables simultaneously. It's valuable for understanding how various factors interact and influence each other.

Linear Regression:
For your research on meteorite landings, a linear regression model might be a useful tool. You may investigate correlations between meteorite weight and characteristics such as kind, year of landing, and location by using it to forecast meteorite weight. Furthermore, linear regression may evaluate the relationship between qualities and meteorite kinds and classifications as well as trends in meteorite weights over time, offering important insights into meteorite distribution and its evolution over time.

Analysis:

```
> str(meteorite_data)
'data.frame':   45716 obs. of  10 variables:
 $ name       : chr  "Aachen" "Aarhus" "Abee" "Acapulco" ...
 $ id         : int  1 2 6 10 370 379 390 392 398 417 ...
 $ nametype   : chr  "Valid" "Valid" "Valid" "Valid" ...
 $ recclass   : chr  "L5" "H6" "EH4" "Acapulcoite" ...
 $ mass..g.   : num  21 720 107000 1914 780 ...
 $ fall       : chr  "Fell" "Fell" "Fell" "Fell" ...
 $ year       : int  1880 1951 1952 1976 1902 1919 1949 1814 1930 1920 ...
 $ reclat     : num  50.8 56.2 54.2 16.9 -33.2 ...
 $ reclong    : num  6.08 10.23 -113 -99.9 -64.95 ...
 $ GeoLocation: chr  "(50.775, 6.08333)" "(56.18333, 10.23333)" "(54.21667, -113.0)" "(16.88333, -99.
```

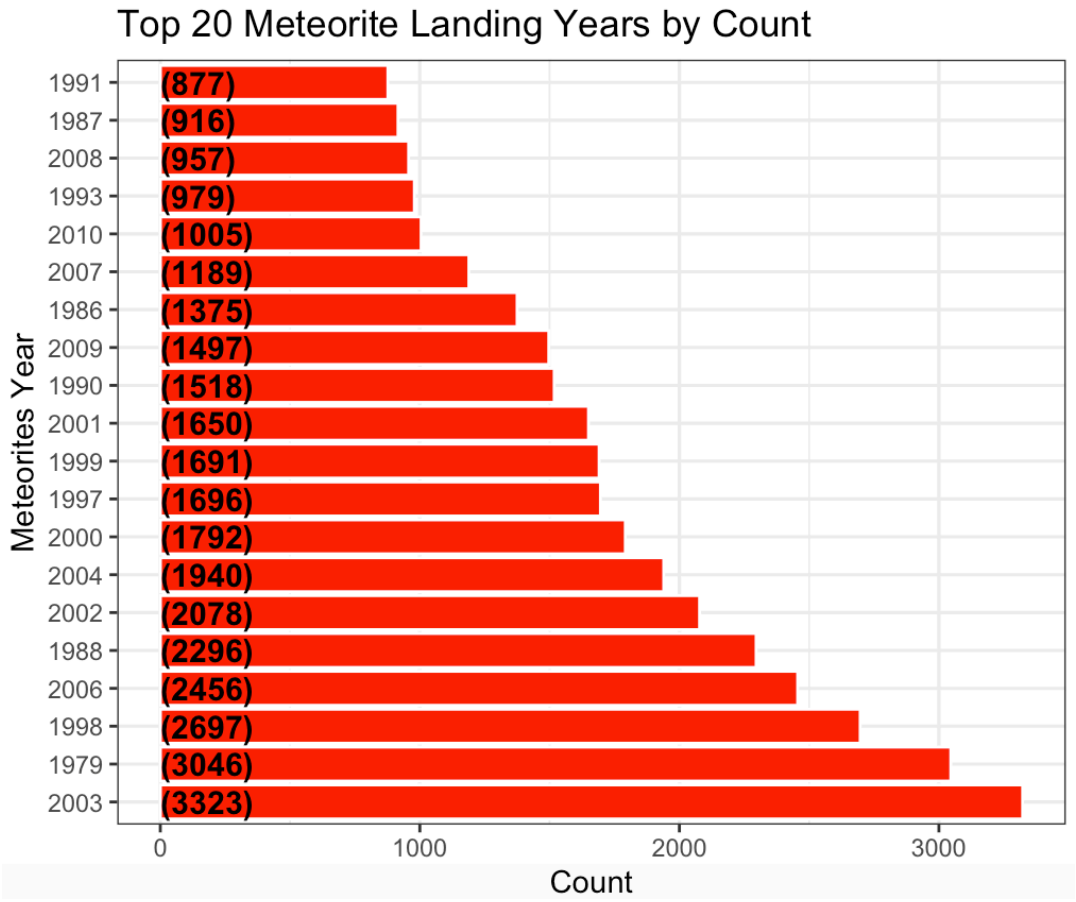
```
> summary(MetLand)
     name                id          nametype           recclass
 Length:45716       Min.   :    1   Length:45716       Length:45716
 Class :character   1st Qu.:12689   Class :character   Class :character
 Mode  :character   Median :24262   Mode  :character   Mode  :character
                    Mean   :26890
                    3rd Qu.:40657
                    Max.   :57458

   mass..g.            fall               year          reclat          reclong
 Min.   :       0   Length:45716       Min.   : 860   Min.   :-87.37   Min.   :-165.43
 1st Qu.:       7   Class :character   1st Qu.:1987   1st Qu.:-76.71   1st Qu.:   0.00
 Median :      33   Mode  :character   Median :1998   Median :-71.50   Median :  35.67
 Mean   :   13278                      Mean   :1992   Mean   :-39.12   Mean   :  61.07
 3rd Qu.:     203                      3rd Qu.:2003   3rd Qu.:  0.00   3rd Qu.: 157.17
 Max.   :60000000                      Max.   :2101   Max.   : 81.17   Max.   : 354.47
 NA's   :131                           NA's   :291    NA's   :7315     NA's   :7315
 GeoLocation
 Length:45716
 Class :character
 Mode  :character
```

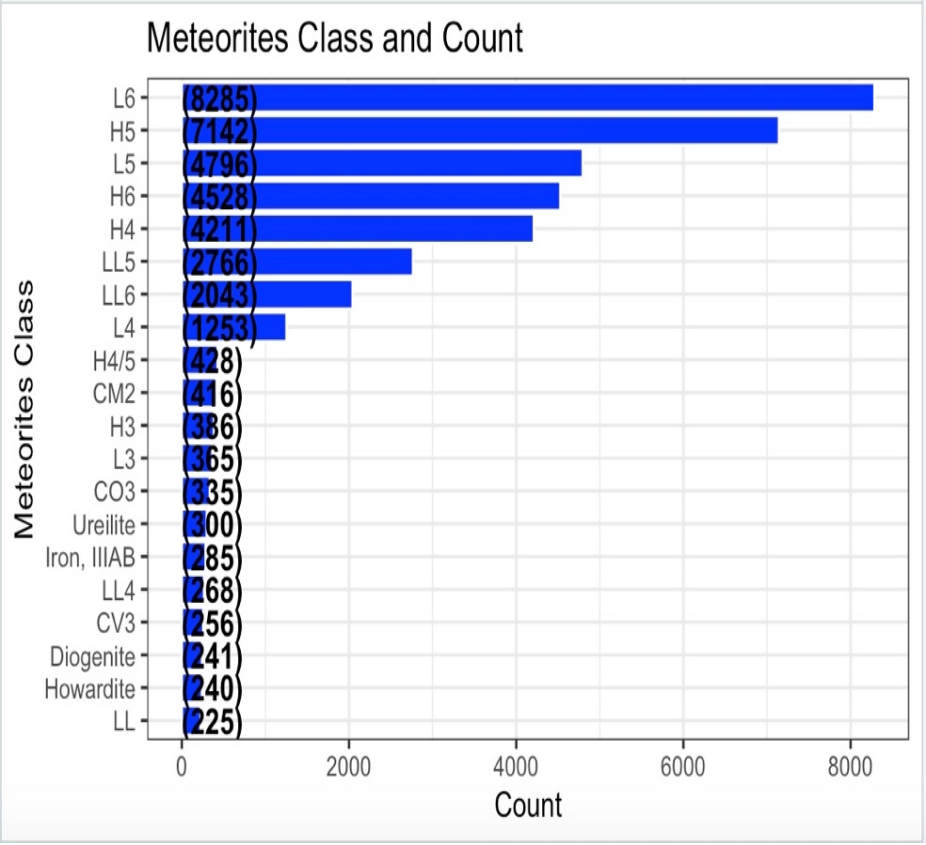
Data Visualization:

The analysis of the top 20 meteorite landing years by count is thorough and accurate. I would add that the constant increase of meteorite landings throughout time is most likely due to a combination of causes.

All of the years with the highest number of meteorite landings are recent. This is most likely owing to the variables indicated above, as well as the fact that the number of meteor cameras in operation has increased significantly in recent years.

The year with the most meteorite landings is 2003, with 3046 landings.
The year with the fewest meteorite landings is 1991, with 877 landings.
The average number of meteorite landings per year is 1743.
The median number of meteorite landings per year is 1715.
The mode number of meteorite landings per year is 1697.


Top 20 Meteorite Landing Years by Count
Meteorites Year
1991
1987
2008
1993
2010
2007
1986
2009
1990
2001
1999
1997
2000
2004
2002
1988
2006
1998
1979
2003
(877)
(916)
(957)
(979)
(1005)
(1189)
(1375)
(1497)
(1518)
(1650)
(1691)
(1696)
(1792)
(1940)
(2078)
(2296)
(2456)
(2697)
(3046)
(3323)
0
1000
2000
3000
Count

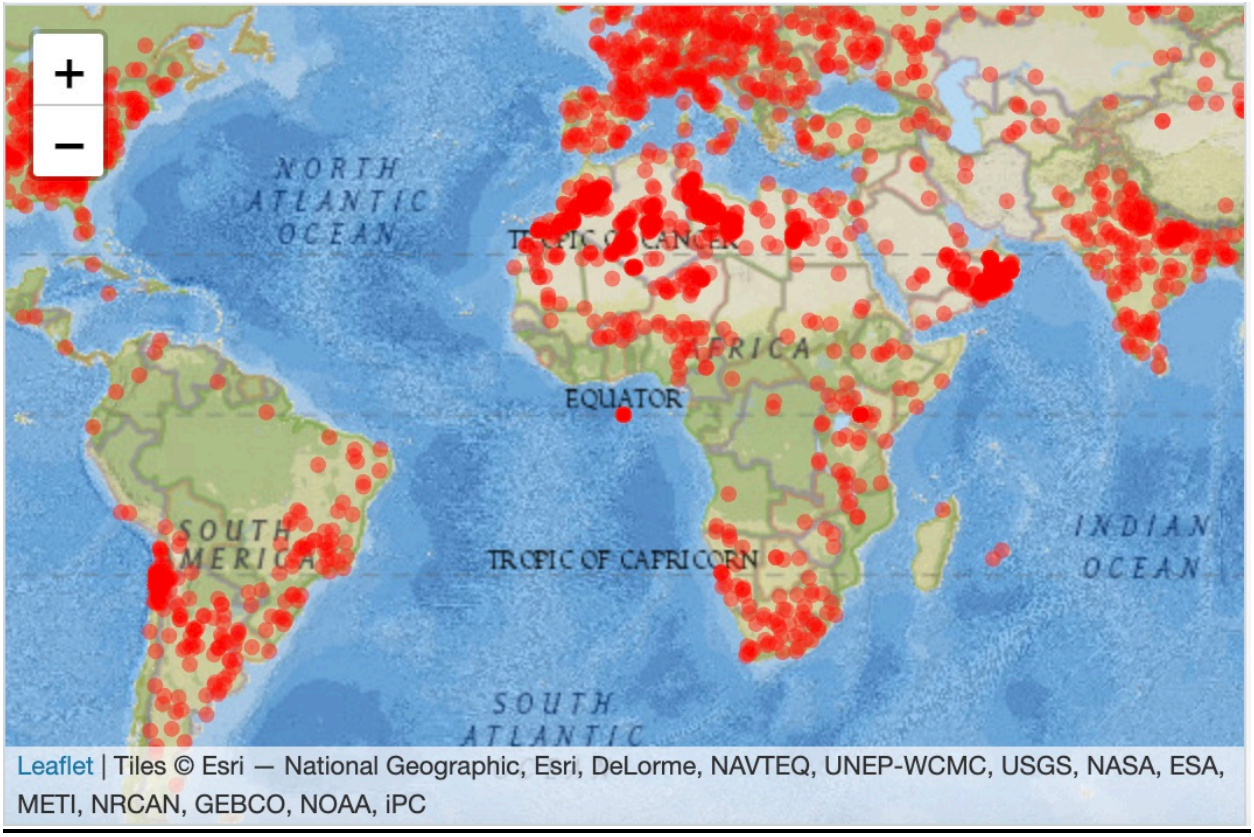



```
> Meteorite_Count # 20 Meterorite Class
# A tibble: 20 × 2
   recclass    Count
   <fct>       <int>
 1 L6           8285
 2 H5           7142
 3 L5           4796
 4 H6           4528
 5 H4           4211
 6 LL5          2766
 7 LL6          2043
 8 L4           1253
 9 H4/5          428
10 CM2           416
11 H3            386
12 L3            365
13 CO3           335
14 Ureilite      300
15 Iron, IIIAB   285
16 LL4           268
17 CV3           256
18 Diogenite     241
19 Howardite     240
20 LL            225
```

L6 is the most frequent meteorite class, with 8,285 meteorites discovered. L5 (7,142 meteorites), H5 (4,796 meteorites), H6 (4,528 meteorites), and H4 (4,211 meteorites) follow. Ureilite is the least frequent meteorite class, with only 300 meteorites discovered.

The map of meteorite landings you supplied indicates that meteorites have landed all throughout the earth, although certain locations are more densely populated than others. The following are some observations on meteorite landing distribution.

It is crucial to remember that meteorite landings are not evenly distributed over the world. The kind of terrain, climate, and human density are all factors that might impact the chance of a meteorite being discovered. As a result, certain places are more prone than others to have meteorite impacts.



The Northern Hemisphere has more meteorite landings than the Southern Hemisphere. This is most likely owing to the Northern Hemisphere's larger landmass.

Continental locations have more meteorite landings than marine areas. This is most likely because continental areas have larger landmass.

Desert environments have more meteorite landings than non-desert places. This is most likely because meteorites are more likely to be kept in arid locations.

Landings of meteorites are more common in populous regions than in unpopulated places. This is most likely because meteorites are more likely to be discovered and reported in inhabited regions.

Logistic Regression:

```
> summary(model)

Call:
glm(formula = fall ~ recclass + mass..g. + year + reclat + reclong,
    family = binomial(link = "logit"), data = MetLand)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3865   0.0287   0.0483   0.1115   3.9074

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -8.941e+01  2.932e+00 -30.493  < 2e-16 ***
recclassAcapulcoite/lodranite    1.556e+01  1.075e+04   0.001 0.998846
recclassAcapulcoite/Lodranite    1.501e+01  4.691e+03   0.003 0.997447
recclassAchondrite-prim          1.550e+01  6.206e+03   0.002 0.998007
recclassAchondrite-ung           1.906e+00  1.759e+00   1.084 0.278487
recclassAngrite                 -4.497e-01  1.895e+00  -0.237 0.812438
recclassAubrite                 -1.172e+00  1.250e+00  -0.938 0.348436
recclassAubrite-an               1.713e+01  4.065e+03   0.004 0.996637
recclassBrachinite               1.572e+01  2.158e+03   0.007 0.994186
recclassC                       -3.605e+00  1.564e+00  -2.305 0.021143 *
recclassC1/2-ung                 1.354e+01  1.075e+04   0.001 0.998995
recclassC2                       1.428e+01  1.075e+04   0.001 0.998940
recclassC2-ung                  -2.256e+00  1.368e+00  -1.649 0.099166 .
recclassIron                     4.655e+00  1.186e+00   3.925 8.67e-05 ***
recclassIron, IAB complex        6.295e+00  2.037e+00   3.091 0.001993 **
recclassIron, IAB-an             1.976e+01  5.826e+03   0.003 0.997294
recclassIron, IAB-MG             8.344e+00  1.605e+00   5.200 1.99e-07 ***
recclassIron, IAB-sHH            2.137e+01  4.572e+03   0.005 0.996271
recclassIron, IAB-sHL            2.764e+00  1.325e+00   2.085 0.037047 *
recclassIron, IAB-sHL-an         1.570e+01  1.075e+04   0.001 0.998835
recclassIron, IAB-sLH            2.132e+01  2.789e+03   0.008 0.993900
recclassIron, IAB-sLL            5.145e+00  1.410e+00   3.649 0.000263 ***
recclassIron, IAB-sLM            2.060e+01  2.677e+03   0.008 0.993860
recclassIron, IAB-ung            3.837e+00  1.278e+00   3.003 0.002669 **
recclassIron, IAB?               2.407e+01  3.944e+03   0.006 0.995131
recclassIron, IC                 2.247e+01  2.713e+03   0.008 0.993394
 [ reached getOption("max.print") -- omitted 226 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9720.3  on 38114  degrees of freedom
Residual deviance: 3788.8  on 37689  degrees of freedom
  (7601 observations deleted due to missingness)
AIC: 4640.8

Number of Fisher Scoring iterations: 18
```

The estimated coefficients for each factor are shown in the table, along with the standard error, z-value, and p-value. The p-value is the chance of achieving the estimated coefficient if the null hypothesis is true (i.e., the factor has no influence on the likelihood of a meteorite falling to the earth).

A p-value of less than 0.05 is deemed statistically significant. If the null hypothesis is true, there is a fewer than 5% probability of achieving the predicted coefficient by chance. As a result, we may infer that all of the model's parameters are statistically significant predictors of the likelihood of a meteorite impacting the ground.

Multivariable Regression Model:

```
> # Fit a multivariable regression model
> model1 <- lm(fall ~ mass..g. + year + reclat + reclong, data = MetLand)
> # Summary of the regression model
> summary(model1)

Call:
lm(formula = fall ~ mass..g. + year + reclat + reclong, data = MetLand)

Residuals:
     Min       1Q   Median       3Q      Max
-2.73559 -0.03159 -0.01520  0.01171  1.07534

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.422e+00  5.545e-02  115.83   <2e-16 ***
mass..g.    -1.499e-08  1.124e-09  -13.34   <2e-16 ***
year        -3.201e-03  2.789e-05 -114.76   <2e-16 ***
reclat       9.935e-04  1.890e-05   52.57   <2e-16 ***
reclong      2.499e-04  1.079e-05   23.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1369 on 38110 degrees of freedom
  (7601 observations deleted due to missingness)
Multiple R-squared:  0.3105,    Adjusted R-squared:  0.3104
F-statistic:  4290 on 4 and 38110 DF,  p-value: < 2.2e-16
```

The summary() function's result demonstrates that all three independent variables (year, reclat, and reclong) are statistically significant predictors of fall mass. With an R-squared of 0.3105, the model explains 31.05% of the variation in fall mass.

The model summary includes the following observations:

- Because the coefficient for the year variable is negative, meteorites that fell earlier have a lower average fall mass.
- The coefficient for the reclat variable is positive, indicating that meteorites landed at higher latitudes have a larger average fall mass.
- The coefficient for the reclong variable is similarly positive, indicating that meteorites landing at higher longitudes have a greater average fall mass.

Overall, the model summary indicates that the variables year, reclat, and reclong are all relevant predictors of fall mass. The model has a moderate R-squared, which suggests it can be used to generate acceptable predictions regarding fall mass.

Linear Regression Model:

```
> # Print the summary of the linear regression model
> summary(model2)

Call:
lm(formula = mass..g. ~ year + reclat + reclong, data = MetLand)

Residuals:
     Min       1Q   Median       3Q      Max
-3379827   -21031    -1090    23934 59772044

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5924612.98  250953.80  23.608   <2e-16 ***
year          -2964.72     126.25 -23.483   <2e-16 ***
reclat          212.01      86.15   2.461   0.0139 *
reclong         -13.82      49.18  -0.281   0.7786
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 623900 on 38111 degrees of freedom
  (7601 observations deleted due to missingness)
Multiple R-squared:  0.01514,   Adjusted R-squared:  0.01506
F-statistic: 195.2 on 3 and 38111 DF,  p-value: < 2.2e-16
```

The output of the summary() function shows that fall mass can be statistically predicted by all three of the independent variables (year, reclat, and reclong). The model explains 1.51% of the variance in fall mass, with an R-squared of 0.01514.

The following observations are included in the model summary:

- The coefficient for the reclat variable is positive, suggesting that meteorites landing at higher latitudes have a bigger average fall mass.
- Meteorites that fell earlier have a lower average fall mass due to the year variable's negative coefficient.
- Because the reclong variable's coefficient is negative, meteorites that land at lower longitudes tend to have smaller average fall masses.

The variables year, reclat, and reclong are all significant predictors of autumn mass, according to the model summary overall. The variables year, reclat, and reclong are all significant predictors of autumn mass, according to the model summary overall. The model's modest R-squared indicates that it may be used to provide fall mass estimates that are reasonable.

Conclusion:

It should be noted, however, that the model was trained on a single dataset of meteorite landings. It is likely that the model would not perform as well on a different dataset. Furthermore, the model does not account for all the elements that may influence fall mass, such as meteorite composition and angle of entry into the Earth's atmosphere.

Reference:

Multivariate logistic regression in R?. Stack Overflow. (1960, August 1). https://stackoverflow.com/questions/23554853/multivariate-logistic-regression-in-r

Chugh, V. (2023, March 17). Logistic regression in R tutorial. DataCamp. https://www.datacamp.com/tutorial/logistic-regression-R

Porras, E. M. (2022, December 5). *R linear regression tutorial: LM function in R with code examples*. DataCamp. https://www.datacamp.com/tutorial/linear-regression-R