**Predicting Sales from Conversations**

**Task A**

|  |  |  |
| --- | --- | --- |
| **Source** | **Target** | **Weight** |
| ES | LS | 2 |
| ES | RX | 2 |
| ES | A8 | 3 |
| ES | A6 | 1.66666667 |
| ES | 3Series | 6 |
| ES | 5Series | 5 |
| ES | 7Series | 2.8 |
| ES | XJ | 1 |
| ES | Sclass | 2.61538462 |
| LS | RX | 1.14285714 |
| LS | A8 | 2.83783784 |
| LS | A6 | 1.28571429 |
| LS | 3Series | 4 |
| LS | 5Series | 3.1 |
| LS | 7Series | 2.75 |
| LS | XJ | 2.88888889 |
| LS | Sclass | 2.67741935 |
| RX | A8 | 1 |
| RX | A6 | 0 |
| RX | 3Series | 4 |
| RX | 5Series | 0 |
| RX | 7Series | 1 |
| RX | XJ | 0 |
| RX | Sclass | 2.5 |
| A8 | A6 | 3.16666667 |
| A8 | 3Series | 5 |
| A8 | 5Series | 5 |
| A8 | 7Series | 3 |
| A8 | XJ | 2.14285714 |
| A8 | Sclass | 2.0625 |
| A6 | 3Series | 5 |
| A6 | 5Series | 5 |
| A6 | 7Series | 5 |
| A6 | XJ | 0 |
| A6 | Sclass | 2.33333333 |
| 3Series | 5Series | 0 |
| 3Series | 7Series | 2 |
| 3Series | XJ | 3 |
| 3Series | Sclass | 3.33333333 |
| 5Series | 7Series | 1 |
| 5Series | XJ | 2 |
| 5Series | Sclass | 2 |
| 7Series | XJ | 3.16666667 |
| 7Series | Sclass | 2.4375 |
| XJ | Sclass | 2.28571429 |
| LS | ES | 1.16666667 |
| RX | ES | 1.66666667 |
| A8 | ES | 3.4 |
| A6 | ES | 6 |
| 3Series | ES | 3.25 |
| 5Series | ES | 4 |
| 7Series | ES | 3.6 |
| XJ | ES | 1 |
| Sclass | ES | 3.28571429 |
| RX | LS | 3.4 |
| A8 | LS | 2.85714286 |
| A6 | LS | 3.33333333 |
| 3Series | LS | 2.8 |
| 5Series | LS | 2.7 |
| 7Series | LS | 2.64285714 |
| XJ | LS | 2.75 |
| Sclass | LS | 2.55405405 |
| A8 | RX | 2.2 |
| A6 | RX | 3.66666667 |
| 3Series | RX | 4 |
| 5Series | RX | 3.6 |
| 7Series | RX | 3.75 |
| XJ | RX | 2 |
| Sclass | RX | 3.63636364 |
| A6 | A8 | 1.375 |
| 3Series | A8 | 3.25 |
| 5Series | A8 | 3.33333333 |
| 7Series | A8 | 2.18181818 |
| XJ | A8 | 2 |
| Sclass | A8 | 2.5 |
| 3Series | A6 | 3 |
| 5Series | A6 | 3 |
| 7Series | A6 | 0 |
| XJ | A6 | 1 |
| Sclass | A6 | 2 |
| 5Series | 3Series | 0 |
| 7Series | 3Series | 5 |
| XJ | 3Series | 0 |
| Sclass | 3Series | 4.33333333 |
| 7Series | 5Series | 2.33333333 |
| XJ | 5Series | 0 |
| Sclass | 5Series | 3 |
| XJ | 7Series | 3.66666667 |
| Sclass | 7Series | 3.11111111 |
| Sclass | XJ | 1.66666667 |

**Table 1: Positive and negative average values of X-Y**

**Task B**

Calculate the **weighted PageRank scores** for each car. What is the correlation between the weighted PageRank and sales figures shown below? A sample python script to calculate weighted PageRank scores is posted on Canvas. You have to install the networkx Python library to make this script work. You also need to fill in all the details from your calculations in Task A.

Also run a regression with sales as the output (dependent variable) and the weighted PageRank score as the predictor (independent) variable. How would you use this approach in the real world to predict, say, the sales of a new model of Tesla? Outline the steps (you don’t have to do an actual analysis).

|  |  |
| --- | --- |
| **Label** | **PageRanks** |
| ES | 0.11075 |
| LS | 0.097626 |
| RX | 0.06191 |
| A8 | 0.121278 |
| A6 | 0.119991 |
| 3Series | 0.111304 |
| 5Series | 0.100445 |
| 7Series | 0.098818 |
| XJ | 0.070036 |
| Sclass | 0.107842 |

**Table 2: Weighted Page Ranks for every Car**

The correlation between weighted page rank and sales is 0.736422. This means that there is a positive correlation between the weighted page rank and sales.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SUMMARY OUTPUT** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ***Regression Statistics*** | |  |  |  |  |  |  |  |
| **Multiple R** | **0.676981511** |  |  |  |  |  |  |  |
| **R Square** | **0.458303966** |  |  |  |  |  |  |  |
| **Adjusted R Square** | **0.347192855** |  |  |  |  |  |  |  |
| **Standard Error** | **72.33530308** |  |  |  |  |  |  |  |
| **Observations** | **10** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **ANOVA** |  |  |  |  |  |  |  |  |
|  | ***df*** | ***SS*** | ***MS*** | ***F*** | ***Significance F*** |  |  |  |
| **Regression** | **1** | **39841.9954** | **39841.9954** | **7.61448384** | **0.0246963** |  |  |  |
| **Residual** | **9** | **47091.5646** | **5232.39607** |  |  |  |  |  |
| **Total** | **10** | **86933.56** |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** | ***Lower 95%*** | ***Upper 95%*** | ***Lower 95.0%*** | ***Upper 95.0%*** |
| **Intercept** | **0** | **#N/A** | **#N/A** | **#N/A** | **#N/A** | **#N/A** | **#N/A** | **#N/A** |
| **X Variable 1** | **620.4014014** | **224.829107** | **2.75943542** | **0.02213258** | **111.802627** | **1129.00018** | **111.802627** | **1129.00018** |

**Table 3 : Regression performed before including luxury input variable**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.751625696 |  |  |  |  |  |  |  |
| R Square | 0.564941187 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.385558835 |  |  |  |  |  |  |  |
| Standard Error | 68.75791903 |  |  |  |  |  |  |  |
| Observations | 10 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 49112.3486 | 24556.1743 | 5.19415923 | 0.0413944 |  |  |  |
| Residual | 8 | 37821.2114 | 4727.65143 |  |  |  |  |  |
| Total | 10 | 86933.56 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 0 | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A | #N/A |
| X Variable 1 | 834.4298197 | 262.741357 | 3.17586021 | 0.01307872 | 228.547164 | 1440.31248 | 228.547164 | 1440.31248 |
| X Variable 2 | -59.18635974 | 42.2664928 | -1.400314 | 0.19898882 | -156.65307 | 38.2803475 | -156.65307 | 38.2803475 |

**Table 4: Regression performed with luxury input variable**

**Analysis for Tesla**

* Scrape data about Tesla reviews or inquiries from car forums like Edmunds.
* Calculate the sentiment scores for reviews of Tesla using sentiment analysis.
* Using sentiment scores calculate average for: X-Y and Y-X.
* Using these average scores calculate the Page Rank.
* We have built a regression model using sales, pagerank and luxury input variable data for 10 models. Now we will use this model to predict the sales figure for Tesla using pagerank and luxury input variable as input data. Luxury input variable value for Tesla is 1, we are assuming.

**Task C**

Task C. For all the links that have a non-zero weight, create a labeled and weighted network using Gephi or NodeXL (works on Windows only). Calculate the PageRank scores ignoring the weights (both tools can calculate these scores). What is the correlation between the PageRank scores and the sales data shown above? Why is this correlation smaller than that between weighted PageRank and sales?

Can you suggest a way to increase the correlation between the PageRank and sales? That is, how can we make the PageRank scores more meaningful? Hint: Make the network more sparse by not showing arrows if the weight is not greater than a minimum value. Make the necessary changes (explain what you did), re-calculate pagerank scores and show the new (higher) correlation.

|  |  |
| --- | --- |
| Id | Pageranks |
| ES | 0.110644 |
| LS | 0.110644 |
| RX | 0.106068 |
| A8 | 0.110644 |
| A6 | 0.085369 |
| 3Series | 0.092524 |
| 5Series | 0.077498 |
| 7Series | 0.109464 |
| XJ | 0.0865 |
| Sclass | 0.110644 |

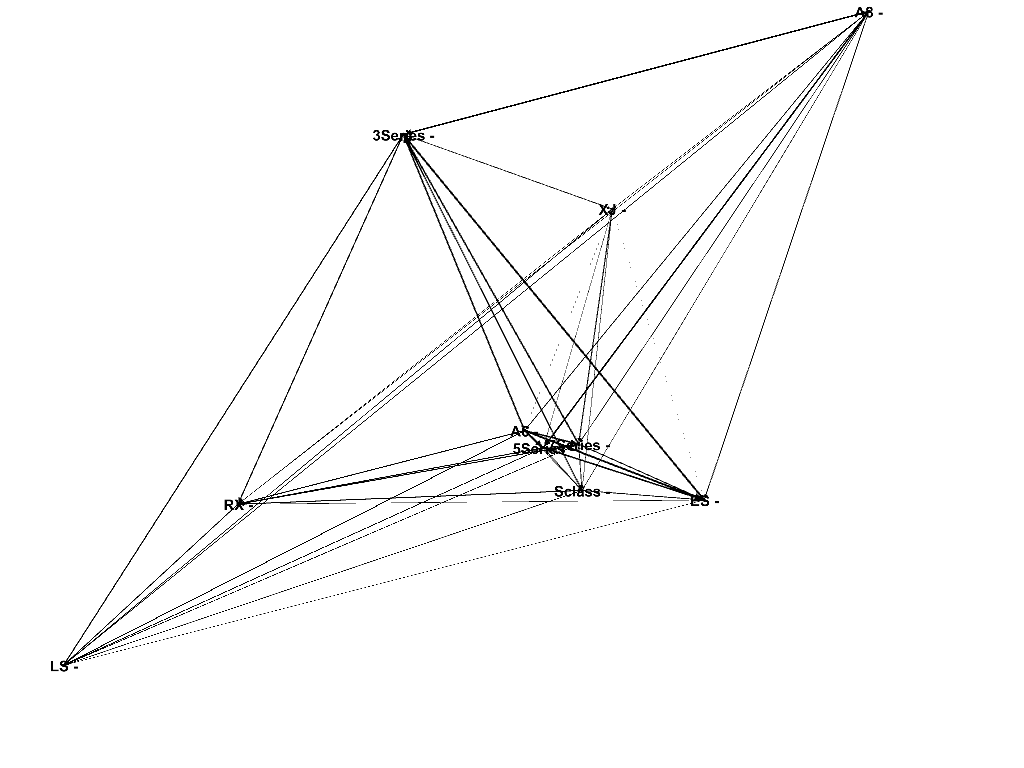
**Table 5: Page Rank scores ignoring the weight**

**Correlation values**

**Weighted = 0.736422**

**Non-weighted = -0.0436**

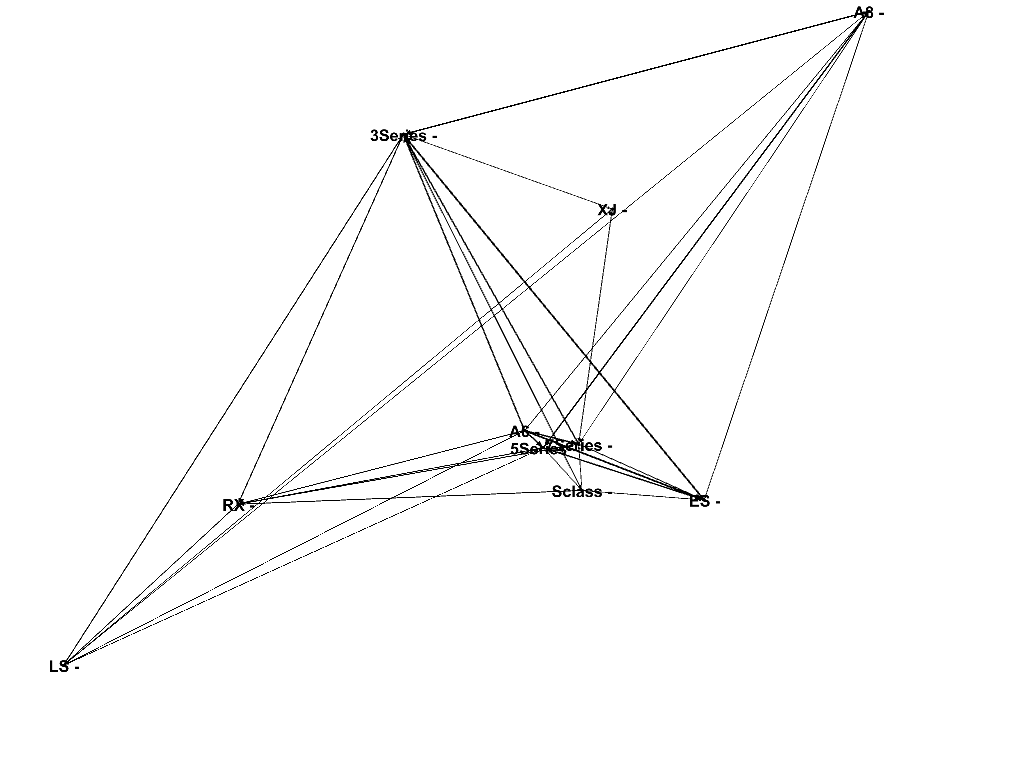
The difference in correlation values between that of weighted PageRank and non-weighted PageRank is because the weighted PageRank takes into consideration the sentiment scores, which in turn increases the correlation ratio. When we calculate the non-weighted PageRank, the value of PageRank is same for many models, but the sales figure are different for different brands. This results in a smaller correlation for the non-weighted PageRank.



**Fig: network graph for weighted pages.**

|  |  |
| --- | --- |
| Id | pageranks |
| ES | 0.116589 |
| LS | 0.09289 |
| RX | 0.104636 |
| A8 | 0.093393 |
| A6 | 0.065659 |
| 3Series | 0.192587 |
| 5Series | 0.096989 |
| 7Series | 0.122445 |
| XJ | 0.075702 |
| Sclass | 0.03911 |

**Table 6 : PageRank for a sparse network**

****

**Fig: Network graph for a sparse network**

We set the threshold edge weight at **2.77** to make the network sparse by removing the connections that have an edge weight less than the threshold. This resulted in an increase in the correlation value to **0.78**. The network should not be made too sparse, so that it does not result in some nodes being disconnected from the network. When the threshold value was increased closer to 3, it resulted in few of the nodes being disconnected. A higher threshold value might result in loss of meaningful data. This is one of the ways to increase the correlation between the PageRank and the sales.