**Group Members:** Christopher Leung, Vinay Burhade, Het Patel, Dhruveen Patel, Manish Kakadiya

**Part 1: The final file after pre-processing and cleaning is shown below.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Average Nurse  Rank | Physician  Social Media  System | Number of Beds | Room Size | Primary Physician  Avg Visits | Average Patient  Stay (days) | Hospital Rank | % Critical Care  Patients | Avg Patient Satisfaction Rate |
| 4 | No | 479 | Private | 6 | 4 | 7 | 31 | 8 |
| 3 | No | 458 | Private | 7 | 4 | 9 | 18 | 11 |
| 2 | No | 356 | Two Beds | 3 | 6 | 9 | 21 | 11 |
| 2 | No | 466 | Two Beds | 6 | 7 | 1 | 23 | 1 |
| 2 | Yes | 467 | Two Beds | 2 | 5 | 5 | 4 | 6 |
| 4 | Yes | 411 | Two Beds | 2 | 7 | 1 | 0 | 1 |
| 3 | Yes | 558 | Two Beds | 9 | 4 | 1 | 38 | 1 |
| 5 | Yes | 238 | More that 2 Beds | 10 | 8 | 9 | 38 | 11 |
| 4 | No | 365 | More that 2 Beds | 3 | 8 | 9 | 39 | 11 |
| 3 | No | 280 | More that 2 Beds | 2 | 8 | 3 | 17 | 4 |
| 1 | No | 477 | More that 2 Beds | 2 | 5 | 5 | 8 | 6 |
| 4 | No | 423 | Private | 10 | 6 | 7 | 5 | 8 |
| 2 | No | 382 | Private | 9 | 3 | 7 | 76 | 8 |
| 5 | No | 339 | Two Beds | 3 | 6 | 3 | 13 | 4 |
| 5 | No | 391 | Two Beds | 7 | 5 | 1 | 12 | 1 |
| 3 | No | 314 | Two Beds | 3 | 5 | 5 | 9 | 6 |
| 4 | No | 238 | Two Beds | 4 | 7 | 3 | 11 | 4 |
| 4 | No | 530 | Two Beds | 3 | 5 | 5 | 23 | 6 |
| 3 | No | 577 | Two Beds | 7 | 5 | 5 | 34 | 6 |
| 5 | Yes | 219 | Private | 7 | 8 | 7 | 17 | 8 |
| 4 | Yes | 474 | More that 2 Beds | 2 | 4 | 1 | 7 | 1 |
| 5 | Yes | 496 | More that 2 Beds | 3 | 3 | 1 | 24 | 1 |
| 4 | Yes | 335 | More that 2 Beds | 2 | 4 | 7 | 37 | 8 |
| 1 | Yes | 545 | More that 2 Beds | 10 | 4 | 9 | 13 | 11 |
| 4 | Yes | 215 | More that 2 Beds | 10 | 6 | 5 | 33 | 6 |
| 2 | Yes | 438 | Two Beds | 9 | 8 | 7 | 14 | 8 |
| 2 | Yes | 579 | Two Beds | 8 | 8 | 1 | 5 | 1 |
| 4 | Yes | 237 | Two Beds | 6 | 4 | 1 | 37 | 1 |
| 4 | Yes | 326 | Two Beds | 3 | 4 | 3 | 3 | 4 |
| 5 | Yes | 449 | Two Beds | 2 | 4 | 3 | 8 | 4 |
| 1 | No | 438 | Private | 2 | 6 | 3 | 17 | 4 |
| 5 | No | 372 | More that 2 Beds | 3 | 7 | 1 | 15 | 1 |
| 4 | No | 585 | More that 2 Beds | 6 | 8 | 3 | 2 | 4 |
| 3 | No | 286 | More that 2 Beds | 4 | 6 | 7 | 36 | 8 |
| 3 | Yes | 419 | More that 2 Beds | 6 | 8 | 3 | 5 | 4 |
| 3 | Yes | 501 | More that 2 Beds | 10 | 5 | 1 | 7 | 1 |
| 4 | Yes | 320 | Two Beds | 5 | 3 | 3 | 20 | 4 |
| 4 | Yes | 401 | Two Beds | 9 | 7 | 7 | 8 | 8 |
| 3 | Yes | 427 | Two Beds | 8 | 7 | 5 | 33 | 6 |
| 3 | Yes | 583 | Two Beds | 8 | 6 | 1 | 6 | 1 |
| 1 | Yes | 280 | Two Beds | 10 | 4 | 7 | 24 | 8 |
| 1 | Yes | 524 | Two Beds | 9 | 6 | 5 | 24 | 6 |
| 5 | Yes | 297 | Two Beds | 2 | 7 | 5 | 7 | 6 |
| 2 | Yes | 537 | Two Beds | 6 | 4 | 7 | 11 | 8 |
| 4 | Yes | 497 | Two Beds | 10 | 5 | 1 | 9 | 1 |
| 2 | Yes | 367 | Two Beds | 4 | 4 | 7 | 17 | 8 |
| 1 | No | 566 | Two Beds | 6 | 4 | 9 | 21 | 11 |
| 3 | No | 230 | More that 2 Beds | 4 | 4 | 7 | 18 | 8 |
| 2 | No | 355 | More that 2 Beds | 5 | 8 | 5 | 1 | 6 |
| 5 | Yes | 293 | More that 2 Beds | 9 | 3 | 9 | 34 | 11 |
| 2 | Yes | 407 | More that 2 Beds | 2 | 5 | 1 | 17 | 1 |
| 1 | Yes | 526 | More that 2 Beds | 9 | 5 | 7 | 20 | 8 |
| 1 | No | 226 | Private | 6 | 6 | 5 | 12 | 6 |
| 1 | No | 408 | Private | 6 | 3 | 5 | 38 | 6 |
| 2 | No | 414 | More that 2 Beds | 5 | 4 | 3 | 17 | 4 |
| 5 | No | 513 | More that 2 Beds | 6 | 6 | 3 | 65 | 4 |
| 4 | No | 221 | Two Beds | 6 | 8 | 1 | 32 | 1 |
| 5 | No | 446 | Two Beds | 9 | 3 | 7 | 39 | 8 |
| 2 | No | 257 | Two Beds | 10 | 8 | 3 | 34 | 4 |
| 3 | No | 469 | Two Beds | 9 | 3 | 5 | 7 | 6 |
| 1 | No | 466 | Two Beds | 6 | 7 | 9 | 34 | 11 |
| 5 | No | 513 | Two Beds | 10 | 7 | 5 | 18 | 6 |
| 1 | Yes | 575 | Two Beds | 10 | 7 | 5 | 17 | 6 |
| 5 | Yes | 597 | Two Beds | 5 | 8 | 10 | 5 | 12 |
| 1 | Yes | 258 | More that 2 Beds | 7 | 5 | 7 | 8 | 8 |
| 1 | Yes | 464 | More that 2 Beds | 7 | 7 | 1 | 9 | 1 |
| 4 | Yes | 294 | Two Beds | 6 | 4 | 7 | 5 | 8 |
| 5 | Yes | 467 | Two Beds | 3 | 7 | 5 | 4 | 6 |
| 5 | No | 508 | Two Beds | 2 | 5 | 3 | 8 | 4 |
| 4 | No | 427 | Two Beds | 10 | 6 | 3 | 35 | 4 |
| 4 | No | 245 | Private | 6 | 7 | 1 | 7 | 1 |

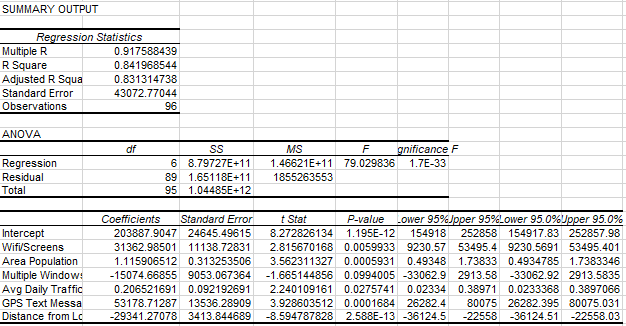
**Tasks carried out on the original file:**

* Column “Hospital region” was removed after filtering out all other regions except “NorthEast” as the analysis is to be focused on just the NorthEast related data.
* Columns “Gender”, “Doctors” and “Nurses” was removed as it did not signify anything and won’t be useful in further analysis. According to the data, the number of doctors and nurses did not affect the final satisfaction rate.
* Column “Hospital Type” was deleted after filtering out the rows for type except “Network” as the analysis is to be focused only on network type of hospitals.
* Column “Average Patient Age” is irrelevant column and therefore are deleted.
* Column “Average Patient Stay (Days)” has two outliers, so these rows are deleted. The information in these cells are ‘2# days and 720 days.
* Column “Average Patient Age” had an average patient age of -33 which is impossible.
* Column “Number of beds” has a field of 1445556 and 20000 which are outliers.
* Column “Hospital Rank” has a rank of 11. The scale is based on 1-10. The rows with data out of the scale were deleted.
* Column “Average nurse rank” has a rank of 6. The scale is based on a 1-5 scale. The rows with data out of the scale were deleted.
* Column “Physical Social Media System” has “ny” as a value. Should only have values Yes or No. Hence, removed.
* Column “Patient Illness”, there is “Obsical” in row 11 that has no meaning. Hence, removed that record.
* Column “% Critical Care Patients”, there are values above 100 which indicated above 100% which is not possible. Hence, these records were removed.

We end up with 9 columns and 72 rows out of 20 columns and 163 rows after carrying out above operations.

This data can be used for further data mining activities like determining the driving variables more accurately and other tasks.

**Part 2: Activities to best manage and operate the fast food restaurant:**

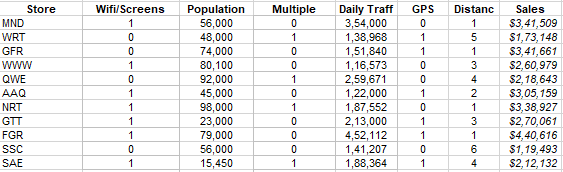
fig. Analysis results of regression in Excel

**Analysis from Regression Statistics Output:**

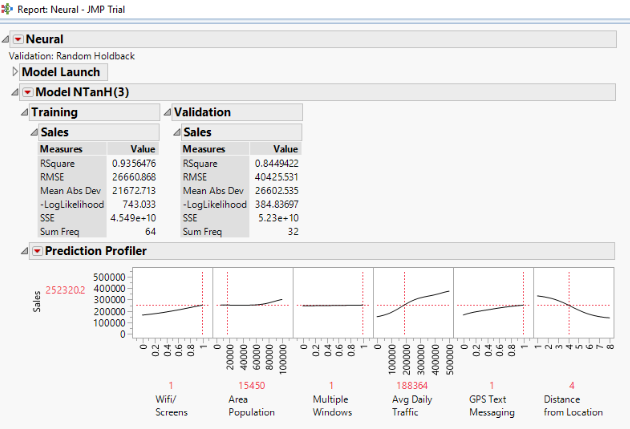
* Multiple R value is 91% which shows the strong positive relationship.
* R^2 is 84% which fits the regression analysis model. In other words, 84% of the dependent variable (y-value) are explained by the independent variables (x-values).
* Wifi/Screens, Area Population, Avg. Daily Traffic, GPS Text Messaging, Distance from Location is either greater than 2 or less than -2 which is likely to be a meaningful addition in our model because changes in the predictor's value are related to changes in the response variable.
* With the increase in Multiple windows, Distance from Location variables it is predicted that sale will decrease as they have negative Coefficients.

**Calculating estimated Sales:**

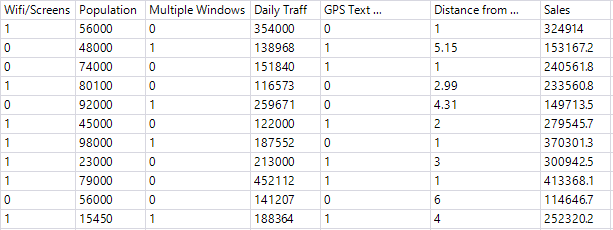
Y=Intercept+(B1\*X1) + (B2\*X2) + ….. + (Bn\*Xn)

To calculate the estimate sales the intercept and coefficients from the Regression Statistics Output are used in the linear regression formula.

**Neural Net methodology Analysis:**



**Prediction of Sales in JMP:**

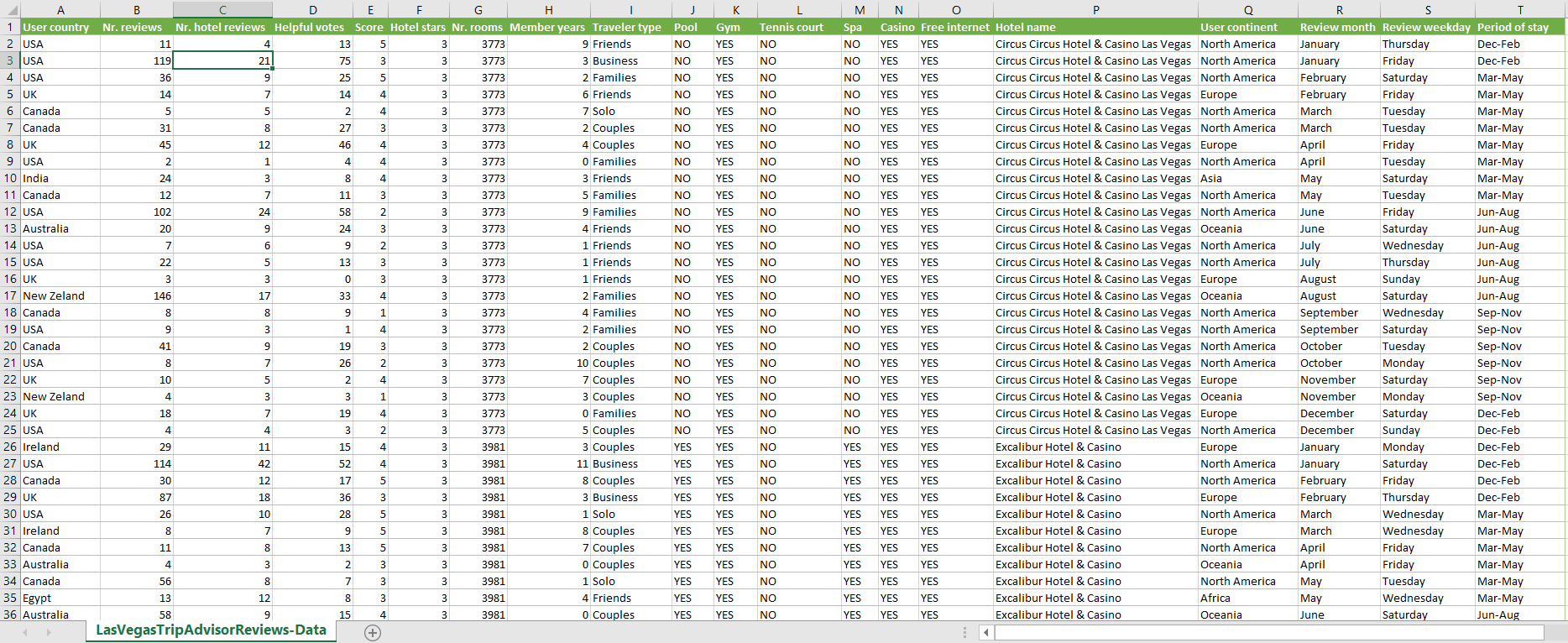


* The neural net model results show the training and validation scores. The R-Square values for both the sets are 0.93 and 0.84 respectively, which shows the model is a good fit for the data.
* The predictive profiler was used to predict the target variable “Sales” for the proposed data. According to the profiler, the variables namely “Population”, “Daily traffic”, “GPS text” and “Distance from location” play the most significant part in predicting the sales. Moreover, the variable “Wifi/Screens” is also significant but not as good predictor as the ones mentioned above. The variable “Multiple Windows” hardly plays any part in driving the target variable.

**Comparison between Regression and Neural Network Model:**

* The R-Square value for the training set of the Neural Net model i.e. 93% is greater than the Regression i.e. 84%. The analysis model of Neural Net is better than the Regression analysis according to the statistical scores.
* The predicted values of the “Sales” target variable of Neural Net model are more accurate than the Regression Analysis.
* Even though there is a difference between the R-Squared values of both the models, the relationship between the independent variables and the target is same for both the models.
* From the analysis it can be determined that the restaurant who gives the GPS texting to the targeted customer and if the customer receive the text message within the ‘1’ mile of distance from food location, the sales of that particular restaurant can increase.
* The area population, Average daily traffic and Wifi/Screens can also affect the sales of the restaurant. In the analysis, the restaurant in the area population above 50,000, having more average daily traffic and also providing free Wifi and have TV Screens can have the high number of sales compared to others.
* Multiple Window is the least significant variable in the analysis.
* In conclusion, a fast food restaurant that has a combination of “GPS Texting”, “Distance from Location”, “Area Population”, “Avg. Daily Traffic” and “Wifi/Screens” with values within the range as mentioned above are more likely to generate more sales and revenue.

**Part 3: Trip Advisor data mining and analysis:**

* The data we selected was Las Vegas Strip Data from Trip Advisor (shown below) between January 2015 to August 2015 and the source website is https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip. The dataset contains 504 rows with 20 total variables. They are User Country, Nr. Reviews, Nr. Hotel Reviews, Helpful Votes, Score, Hotel Stars, Nr. Rooms, Member years, Traveller Type, Pool, Gym, Tennis Court, Spa, Casino, Free Internet, Hotel Name, User Continent, Review Month, Review Weekday, and Period of Stay.
* We felt this data was meaningful and could create a story because the data represents the users on trip advisor rating their Las Vegas trip on a scale (1 to 5) on how their experience was. The other variables are potential driver variables that would determine the overall rating such as Hotel stars or Helpful votes or if having a pool makes a difference.
* Variable Description:
  + User Country = Country where User is from
  + Nr. Reviews = Number of reviews for trip advisor
  + Nr. Hotel Reviews = Number of Hotel Reviews
  + Helpful Votes = Number of Helpful Votes for each review. This is a “like” button for the review, meaning this is number of users who liked the review
  + Score = The Trip Advisor Score ranging 1 to 5. This is the overall score of the entire trip including hotel.
  + Hotel stars = Number of Stars for the Hotel
  + Nr. Rooms = Number of Rooms
  + Member Years = Number of Years the Member has been with Trip Advisor
  + Traveller Type = The user can either be or be with (Solo, Friends, Business, Couples or Families)
  + Pool = Is there a pool? (Yes or No)
  + Gym = Is there a gym? (Yes or No)
  + Tennis Court = Is there a tennis court? (Yes or No)
  + Spa = Is there a Spa? (Yes or No)
  + Casino = Is there a Casino? (Yes or No)
  + Internet = Is there free internet? (Yes or No)
  + Hotel name = The name of Hotel
  + User Continent = Continent where user is from
  + Review Month = The month of user review
  + Review Weekday = The weekday in which user reviewed (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday)
  + Period of Stay = The months the user stayed in vacation in Las Vegas (Dec-Feb, Mar-May, Jun-Aug, Sep-Nov)
* **In order to mine the data:**

We had to remove variables Nr. Rooms (column G), Member Years (column H), User Continent (column Q), Review month (column R), and Review weekday (column S) because there are missing values or blank values in them. In addition, we excluded Hotel name (column P) and Period of stay (column T) because they provided nothing significant as a driver variable.

* **Target Variables and Driver Variables:**

We decided the target variable would be the score of the trip on trip advisor (column E). This is also the best indicator to rate the trip experience on whether it was a 5 (highest rating) or 1 (lowest rating). The driver variables we used were Nr. Reviews, Nr. Hotel reviews, Helpful votes, Traveller type, Pool, Gym, Tennis Court, Spa, Casino, Free Internet, Hotel Stars, and User Country.

* **Data Transformations?**

We decided not to use data transformations even when columns H to M only use Yes or No values. When we did transform the Yes/No to binary (0/1), the neural network model did not improve with such transformation, so this would be unnecessary. Also, the variable Period of Stay (column T) did not help in developing any sort of patterns when we ran neural network considering this as one of the independent variables. There are not enough categories to determine if seasonality had an effect on our data mining analysis. In order to see such an effect, we would need previous seasonal data – at least 1-2 years’ worth of data. The other driver variables had numerical values except for User Country. We can toggle through different countries for checking how users from different countries affect the rating. It would help to interpret the user country’s relationship to the target variable (score).

* **Enough data?**

After reviewing the amount of data which was 504 rows of data, it did appear a bit small for our data mining project. However, the dataset also contained 20 variables comprised of 1 target and 19 drivers. Because of this, we thought the data set would be good for analysis and given the fact freely available good data would be hard to scrap on internet, we felt this would be a great data mining project to analyse.

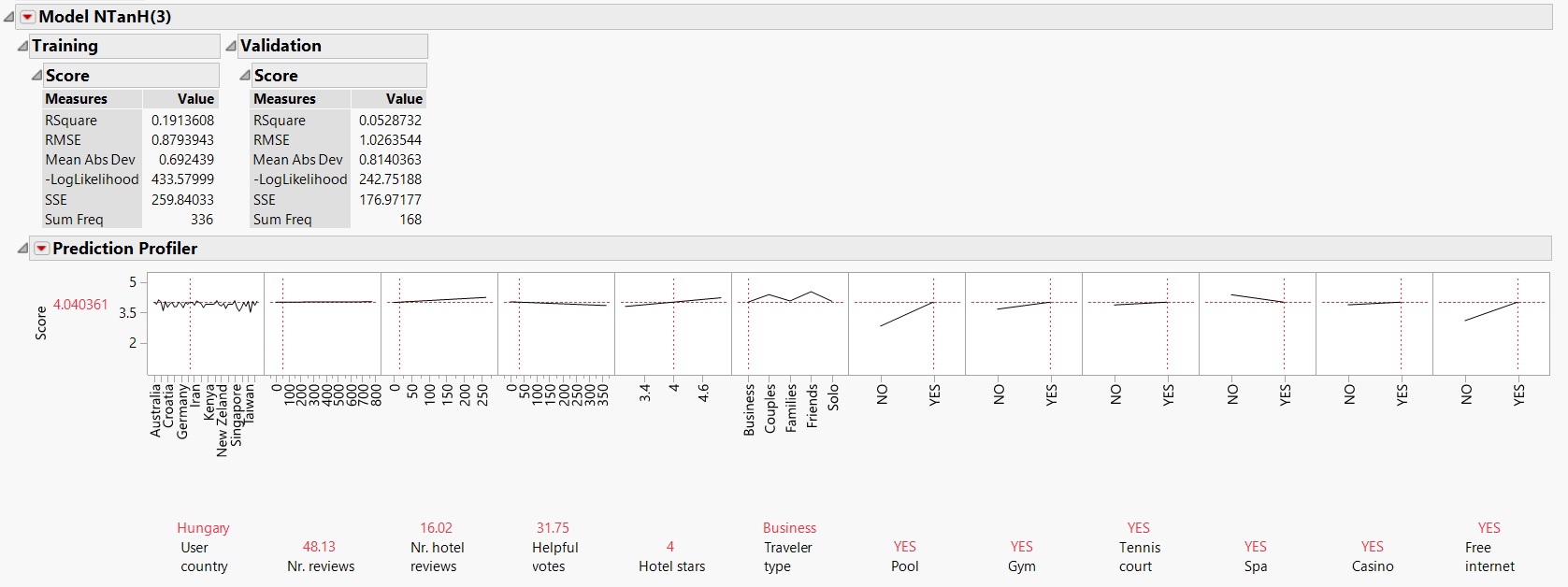
* **Multicollinearity?**

We did not use regression in our data mining project and multicollinearity will not be an issue because of this.

* **Outliers?**

After examining the data set there were a few outliers but to run our model it will work fine as the histograms follow a lognormal distribution (heavy closer to 0) and since the outliers aren’t huge in proportion, we can keep the data set as it is.

* **Data Mining**

With our data set we decided to use neural networks using target variable Score (column E) and driver variables Nr. Reviews, Nr. Hotel reviews, Helpful votes, Traveller type, Pool, Gym, Tennis Court, Spa, Casino, Free Internet, Hotel Stars, and User Country. The neural network is generated below with holdback proportion of 33%, meaning 67% of data is training and 33% of data is for testing how accurate the model is.

The R-Square measures how well the model explains the variation in the target variable. In this case, for the training data set it is 19% and for validation test data set it is 5%. The model is therefore not very accurate, but most models aren’t with real data set. The RMSE is Root Mean Square Error which measures the standard deviation of residuals (or prediction errors). The errors are relatively the same which means the model has good precision, it can reproduce almost similar amount of prediction error. Since RMSE for validation is slightly higher than training that means there may be some overfitting but very small as this is maybe 0.1 to 0.2 off. Mean Absolute Deviation is the average prediction error per observation and their difference is similar to RMSE’s difference, it is also acknowledging that there may be some overfitting. Log likelihood is used to maximize the optimal values of the estimated coefficients. The Sum of Squared Errors of prediction (SSE) is the sum of squares of residuals (deviations predicted from actual empirical values of data), basically deviations of the prediction to the actual values. We believe it is by random nature that the validation SSE is higher than training SSE. By some chance the validation model fit had less deviations than the training model. But this could be because of the random 33% sample used for validation by the model was closer to the predicted values. The Sum Freq is the number of rows or the split applied to original data set.

As you can see in the profiler, User Country and Nr. Reviews does not have much of a relationship with the target variable. The next driver variable Nr. Hotel reviews has a slight positive relationship with score, but we anticipated that it would be a larger positive slope. We believe the reason for this is that Nr. Hotel reviews can contain positive and negative reviews. Both positive and negative Hotel reviews can counter each other’s effects and balance the score out, hence the lack of a higher positive relationship between Number Hotel reviews and Score. The second hypothesis would be that Score is not dependent on Hotel’s rating. But looking at Hotel Stars variable, you can tell that there is a positive relationship. Therefore, the second hypothesis would not make sense as there is a positive relationship between Score and Number Hotel Reviews.

The Helpful Votes variable has a negative relationship with Score, but this could be because the number of Helpful votes favour the more negative Number of reviews. The next variable Traveller Type has two categories Couples and Friends that boost up the target variable Score, which may indicate that couples and friends may have a better, positive experience than those who have responsibility. Such as, the other categories (Business, Families or Solo) scored less because they don’t enjoy the vacation as much because of work related responsibilities or family responsibilities or going solo makes the experience not as great as you would want to have fun with friends.

The next variable Pool indicates a positive relationship with score and would make perfect sense as having a pool in a hotel should be a standard, one without it would receive a lower overall score. There is also a slight positive relationship with variables Gym, Tennis Court with the target variable indicating it should be part of a hotel or else the trip will have a lower score. Surprisingly, the spa variable has a negative relationship with score, this could be anomalous because even if you don’t pay for a spa, it’s a great amenity to have in a hotel. The next surprising variable is Casino as it has a slight positive relationship, we would imagine this to be higher since Las Vegas is known for being a huge casino gambling city – We think this might be anomalous also. A big booster for score is variable Free Internet which has a large positive relationship with the target variable. It makes sense as people are more glued to their cell phones and laptops that free internet is a mandatory amenity to have, to which it increases score.

To conclude, the driver variables that affect the target variable is Number of Hotel Reviews, Hotel Stars, Traveller Type (Couples and Friends), and having all hotel amenities including spa even if it does decrease score because we believe this was anomalous. A hotel in Las Vegas may have its ratings(score) improved by keeping in mind the combination of all these driver variables with values that positively affect the target variable.