**Data Analysis for Neonates**

Understanding the analysis of correlation between the fields like Source, Year, APRDRG, Birthweights with the average length of stay.

Data is prepared and average length of stay is calculated based on different variables. Plots are created based on the data to understand the correlation between different variables.

**Data Preparation**

For Jupyter Notebook, the data is converted in csv from excel and loaded as dataframe. Null values are dropped to clean the data. Then the data in data frame is grouped by SOURCE, YEAR, APRDRG\_VER20 and average length of stay is calculated.

Exhibit1: The table consist of data group by Source, Year, APRDRG\_VER20 and the average length of stay.

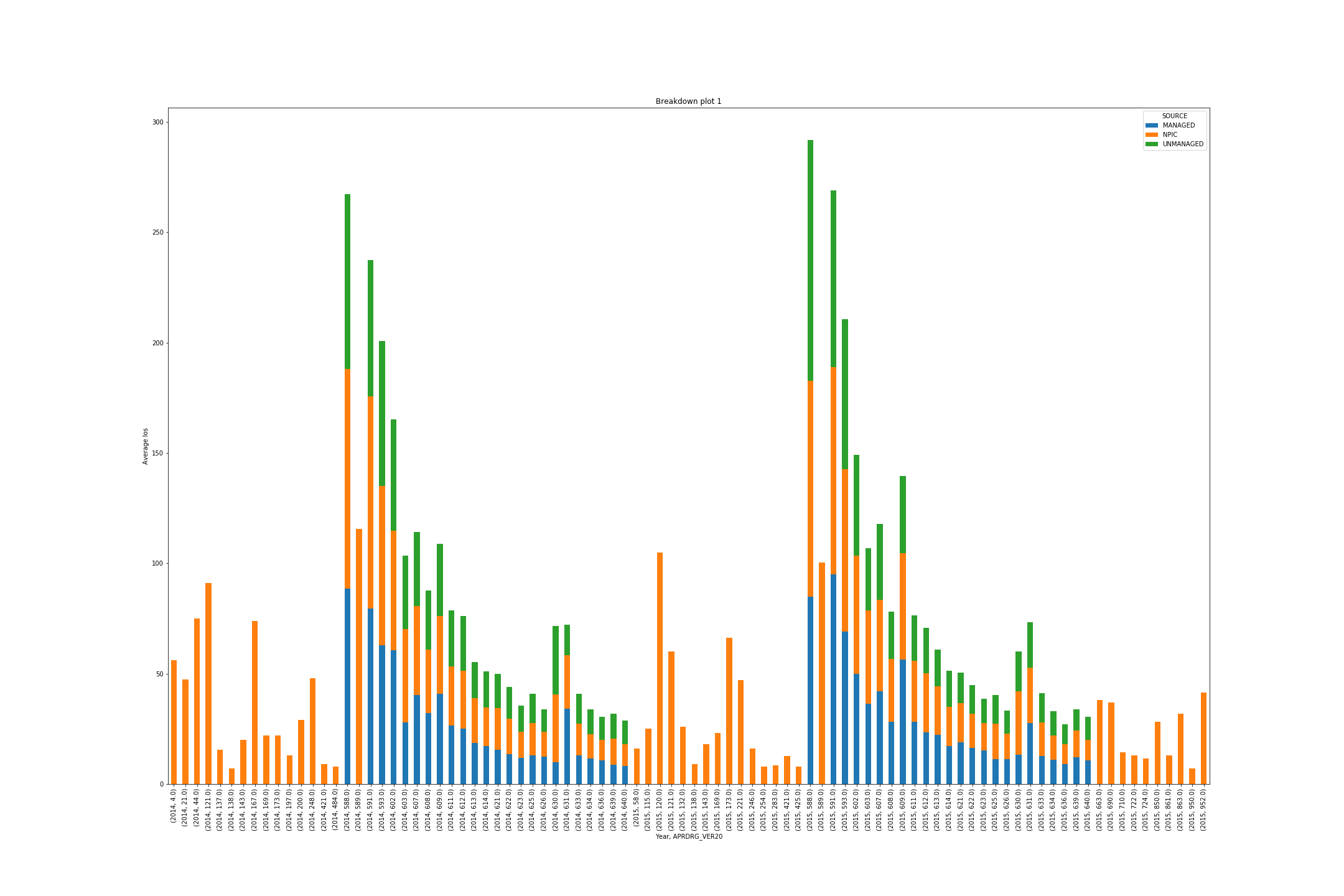
**The derived table from the Panda**

|  |  |  |  |
| --- | --- | --- | --- |
| **YEAR** | **SOURCE** | **APRDRG\_VER20** | **Total Average Los** |
| **2014** | **NPIC** | **4** | **56** |
| **NPIC** | **21** | **47.5** |
| **NPIC** | **44** | **75** |
| **NPIC** | **121** | **91** |
| **NPIC** | **137** | **15.5** |
| **NPIC** | **138** | **7** |
| **NPIC** | **143** | **20** |
| **NPIC** | **167** | **74** |
| **NPIC** | **169** | **22** |
| **NPIC** | **173** | **22** |
| **NPIC** | **197** | **13** |
| **NPIC** | **200** | **29** |
| **NPIC** | **248** | **48** |
| **NPIC** | **421** | **9** |
| **NPIC** | **484** | **8** |
| **MANAGED** | **588** | **88.62068966** |
| **NPIC** | **99.38095238** |
| **UNMANAGED** | **79.25** |
| **NPIC** | **589** | **115.5714286** |
| **MANAGED** | **591** | **79.5** |
| **NPIC** | **96.2972973** |
| **UNMANAGED** | **61.75** |
| **MANAGED** | **593** | **62.96296296** |
| **NPIC** | **71.97964377** |
| **UNMANAGED** | **65.67857143** |
| **MANAGED** | **602** | **60.66666667** |
| **NPIC** | **54.08809524** |
| **UNMANAGED** | **50.45454545** |
| **MANAGED** | **603** | **27.83333333** |
| **NPIC** | **42.35403727** |
| **UNMANAGED** | **33.42857143** |
| **MANAGED** | **607** | **40.22413793** |
| **NPIC** | **40.31685393** |
| **UNMANAGED** | **33.54** |
| **MANAGED** | **608** | **32.1** |
| **NPIC** | **28.9025974** |
| **UNMANAGED** | **26.57142857** |
| **MANAGED** | **609** | **40.76190476** |
| **NPIC** | **35.30337079** |
| **UNMANAGED** | **32.77272727** |
| **MANAGED** | **611** | **26.45238095** |
| **NPIC** | **26.82899628** |
| **UNMANAGED** | **25.3125** |
| **MANAGED** | **612** | **24.98591549** |
| **NPIC** | **26.27620397** |
| **UNMANAGED** | **24.99115044** |
| **MANAGED** | **613** | **18.70588235** |
| **NPIC** | **20.2688172** |
| **UNMANAGED** | **16.15** |
| **MANAGED** | **614** | **17.19354839** |
| **NPIC** | **17.37792103** |
| **UNMANAGED** | **16.50588235** |
| **MANAGED** | **621** | **15.48484848** |
| **NPIC** | **18.87709497** |
| **UNMANAGED** | **15.5952381** |
| **MANAGED** | **622** | **13.56** |
| **NPIC** | **16.18120805** |
| **UNMANAGED** | **14.36448598** |
| **MANAGED** | **623** | **11.75** |
| **NPIC** | **12.06857143** |
| **UNMANAGED** | **11.72** |
| **MANAGED** | **625** | **12.87179487** |
| **NPIC** | **14.74449339** |
| **UNMANAGED** | **13.3015873** |
| **MANAGED** | **626** | **12.30769231** |
| **NPIC** | **11.36204576** |
| **UNMANAGED** | **10.02702703** |
| **MANAGED** | **630** | **10** |
| **NPIC** | **30.51515152** |
| **UNMANAGED** | **31** |
| **MANAGED** | **631** | **34.125** |
| **NPIC** | **24.15286624** |
| **UNMANAGED** | **14** |
| **MANAGED** | **633** | **12.875** |
| **NPIC** | **14.55657492** |
| **UNMANAGED** | **13.40625** |
| **MANAGED** | **634** | **11.52427184** |
| **NPIC** | **11.15943728** |
| **UNMANAGED** | **11.10958904** |
| **MANAGED** | **636** | **10.73076923** |
| **NPIC** | **9.362068966** |
| **UNMANAGED** | **10.42857143** |
| **MANAGED** | **639** | **8.8** |
| **NPIC** | **11.83225806** |
| **UNMANAGED** | **11.2173913** |
| **MANAGED** | **640** | **8.181818182** |
| **NPIC** | **10** |
| **UNMANAGED** | **10.47058824** |
| **2015** | **NPIC** | **58** | **16** |
| **NPIC** | **115** | **25** |
| **NPIC** | **120** | **105** |
| **NPIC** | **121** | **60** |
| **NPIC** | **132** | **26** |
| **NPIC** | **138** | **9** |
| **NPIC** | **143** | **18.16666667** |
| **NPIC** | **169** | **23** |
| **NPIC** | **173** | **66.33333333** |
| **NPIC** | **221** | **47.2** |
| **NPIC** | **246** | **16** |
| **NPIC** | **254** | **8** |
| **NPIC** | **283** | **8.5** |
| **NPIC** | **421** | **12.66666667** |
| **NPIC** | **425** | **8** |
| **MANAGED** | **588** | **85** |
| **NPIC** | **97.74626866** |
| **UNMANAGED** | **109.2** |
| **NPIC** | **589** | **100.5** |
| **MANAGED** | **591** | **94.94444444** |
| **NPIC** | **93.97142857** |
| **UNMANAGED** | **80** |
| **MANAGED** | **593** | **69.15217391** |
| **NPIC** | **73.37844612** |
| **UNMANAGED** | **68.15384615** |
| **MANAGED** | **602** | **49.84090909** |
| **NPIC** | **53.69786096** |
| **UNMANAGED** | **45.61290323** |
| **MANAGED** | **603** | **36.25** |
| **NPIC** | **42.37888199** |
| **UNMANAGED** | **28.25** |
| **MANAGED** | **607** | **41.97058824** |
| **NPIC** | **41.37731481** |
| **UNMANAGED** | **34.38095238** |
| **MANAGED** | **608** | **28.05882353** |
| **NPIC** | **28.65325077** |
| **UNMANAGED** | **21.36363636** |
| **MANAGED** | **609** | **56.42857143** |
| **NPIC** | **48.1625** |
| **UNMANAGED** | **34.90909091** |
| **MANAGED** | **611** | **28.07017544** |
| **NPIC** | **27.84782609** |
| **UNMANAGED** | **20.56603774** |
| **MANAGED** | **612** | **23.33009709** |
| **NPIC** | **26.97481481** |
| **UNMANAGED** | **20.57142857** |
| **MANAGED** | **613** | **22.38461538** |
| **NPIC** | **21.87640449** |
| **UNMANAGED** | **16.61111111** |
| **MANAGED** | **614** | **17.1969697** |
| **NPIC** | **17.83185841** |
| **UNMANAGED** | **16.40322581** |
| **MANAGED** | **621** | **18.8** |
| **NPIC** | **17.85294118** |
| **UNMANAGED** | **13.81578947** |
| **MANAGED** | **622** | **16.2826087** |
| **NPIC** | **15.52554745** |
| **UNMANAGED** | **13.03703704** |
| **MANAGED** | **623** | **15.3125** |
| **NPIC** | **12.40677966** |
| **UNMANAGED** | **10.92307692** |
| **MANAGED** | **625** | **11.35135135** |
| **NPIC** | **15.9049676** |
| **UNMANAGED** | **13.01923077** |
| **MANAGED** | **626** | **11.28** |
| **NPIC** | **11.48584906** |
| **UNMANAGED** | **10.54166667** |
| **MANAGED** | **630** | **13.25** |
| **NPIC** | **28.7037037** |
| **UNMANAGED** | **18** |
| **MANAGED** | **631** | **27.70588235** |
| **NPIC** | **24.88953488** |
| **UNMANAGED** | **20.75** |
| **MANAGED** | **633** | **12.63636364** |
| **NPIC** | **15.2925** |
| **UNMANAGED** | **13.16666667** |
| **MANAGED** | **634** | **11.10227273** |
| **NPIC** | **11.01705238** |
| **UNMANAGED** | **10.79166667** |
| **MANAGED** | **636** | **9.076923077** |
| **NPIC** | **8.921195652** |
| **UNMANAGED** | **9.2** |
| **MANAGED** | **639** | **12.17857143** |
| **NPIC** | **11.98984772** |
| **UNMANAGED** | **9.8** |
| **MANAGED** | **640** | **10.75** |
| **NPIC** | **9.404145078** |
| **UNMANAGED** | **10.29411765** |
| **NPIC** | **663** | **38** |
| **NPIC** | **690** | **37** |
| **NPIC** | **710** | **14.5** |
| **NPIC** | **722** | **13** |
| **NPIC** | **724** | **11.5** |
| **NPIC** | **850** | **28.125** |
| **NPIC** | **861** | **13** |
| **NPIC** | **863** | **31.82758621** |
| **NPIC** | **950** | **7** |
| **NPIC** | **952** | **41.57142857** |



**The pivot table by Excel**

**The plot for the breakdown of the average length of stay vs Source, Year and APRDRG**

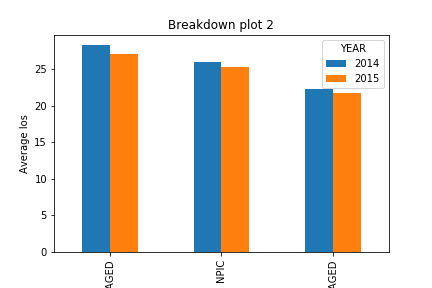
****

To check the relation between Sources and average length of stay with time (Years), the variable APRDRG\_VER20 is disregarded and the average of the length of stay is calculated.

The table is shown below.

|  |  |  |
| --- | --- | --- |
| **SOURCE** | **YEAR** | **Total Average Los** |
| **MANAGED** | **2014** | 28.36192714 |
| **2015** | 27.10942579 |
| **NPIC** | **2014** | 26.05881045 |
| **2015** | 25.36999793 |
| **UNMANAGED** | **2014** | 22.3098434 |
| **2015** | 21.82361111 |

The plot to show the aggregated average of length of stay Vs Source and Year.



**Conclusion**

The plot shows that there is trend in the correlation of the APRDRG vs average length of stay for each Sources.

The first plot also reflects that the length of stay increases with time for the three sources from 2014 to 2015.

The second plot shows managed cases have highest average length of stay for both the year, then NPIC sources have the higher average length of stay than unmanaged cases for both the year.

**Analysis of the distribution of managed cases by APR DRG**

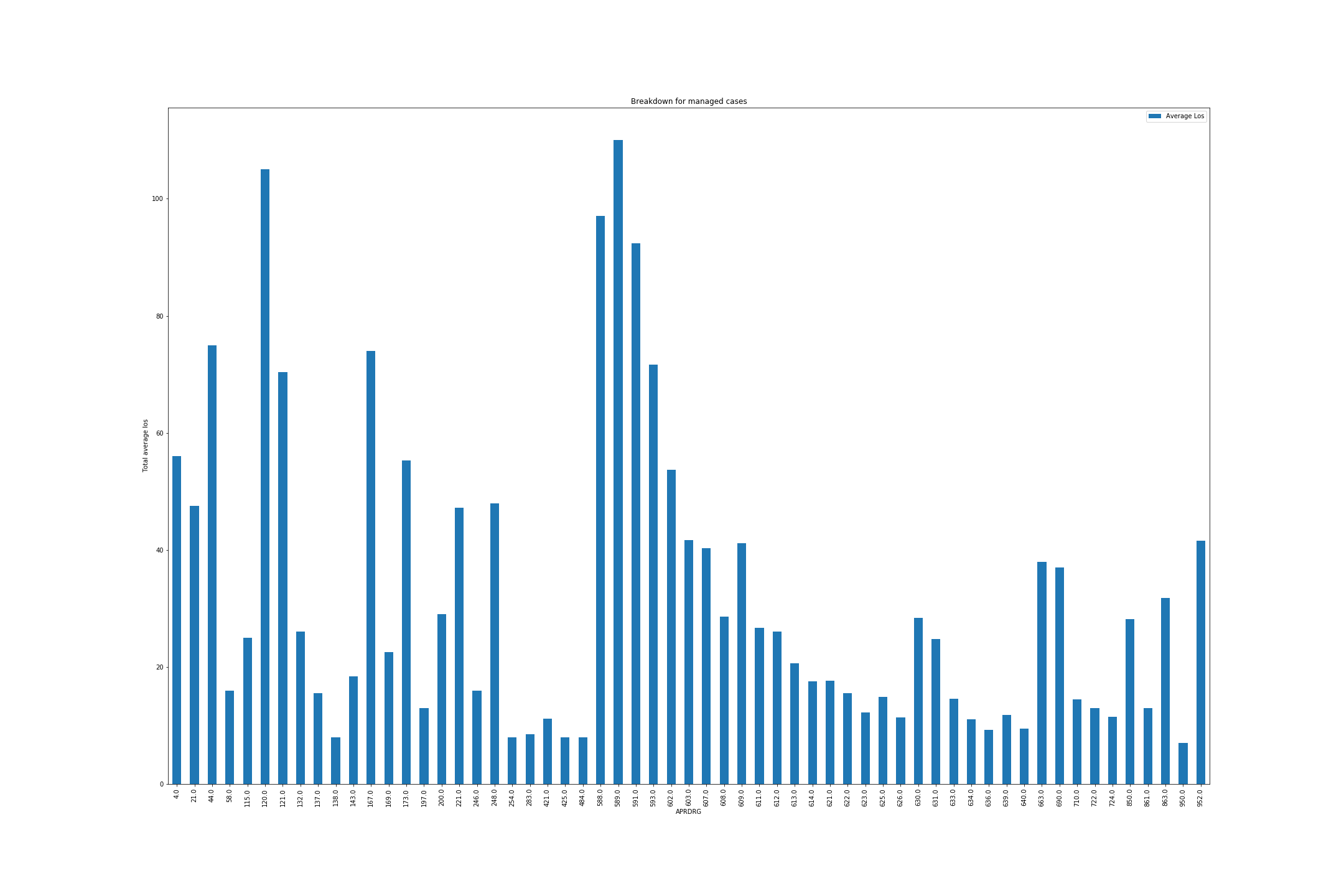
**Data Preparation**

The data is filtered to obtain the cases for only Managed Source. Then the data is aggregated based on APRDRG\_VER20 for both the year 2014 and 2015 and the average length of stay is calculated.

The table showing Average los for APRDRG for only managed case.

|  |  |
| --- | --- |
| **APRDRG\_VER20** | **Average Los** |
| **4** | **56** |
| **21** | **47.5** |
| **44** | **75** |
| **58** | **16** |
| **115** | **25** |
| **120** | **105** |
| **121** | **70.33333333** |
| **132** | **26** |
| **137** | **15.5** |
| **138** | **8** |
| **143** | **18.42857143** |
| **167** | **74** |
| **169** | **22.5** |
| **173** | **55.25** |
| **197** | **13** |
| **200** | **29** |
| **221** | **47.2** |
| **246** | **16** |
| **248** | **48** |
| **254** | **8** |
| **283** | **8.5** |
| **421** | **11.2** |
| **425** | **8** |
| **484** | **8** |
| **588** | **97.02505695** |
| **589** | **110.0909091** |
| **591** | **92.36027714** |
| **593** | **71.62579281** |
| **602** | **53.66736842** |
| **603** | **41.63112392** |
| **607** | **40.29589041** |
| **608** | **28.62028986** |
| **609** | **41.12236287** |
| **611** | **26.73691275** |
| **612** | **26.00690449** |
| **613** | **20.59027778** |
| **614** | **17.52519029** |
| **621** | **17.64595104** |
| **622** | **15.51701134** |
| **623** | **12.22528736** |
| **625** | **14.88718412** |
| **626** | **11.38448176** |
| **630** | **28.39361702** |
| **631** | **24.82585752** |
| **633** | **14.58473625** |
| **634** | **11.10624371** |
| **636** | **9.20959596** |
| **639** | **11.8021164** |
| **640** | **9.507890961** |
| **663** | **38** |
| **690** | **37** |
| **710** | **14.5** |
| **722** | **13** |
| **724** | **11.5** |
| **850** | **28.125** |
| **861** | **13** |
| **863** | **31.82758621** |
| **950** | **7** |
| **952** | **41.57142857** |

The plot for APRDRG vs average length of Stay for managed cases.



**Conclusion**

For managed cases the plot shows that for different APRDRG, the total average length of stay is different.

It also reveals that code 589 for APRDRG has highest total average length of stay and code 959 APRDRG has lowest total average length of stay.

**Exhibit 2 - A table that summarizes the data by the following birthweights.**

**Results should be shown separately by sources, but the two years can be combined.**

**Data Preparation**

A field Birth Weight Range is created to arrange the data based on the Birthweights. The data is aggregated to calculate the total average length of stay based on Birth Weight Ranges and APRDRG\_VER20. Below is list of Birth Weight Ranges

< 750 grams

-          750 - 1000 grams

-          1000 - 1249 grams

-          1250 - 1499 grams

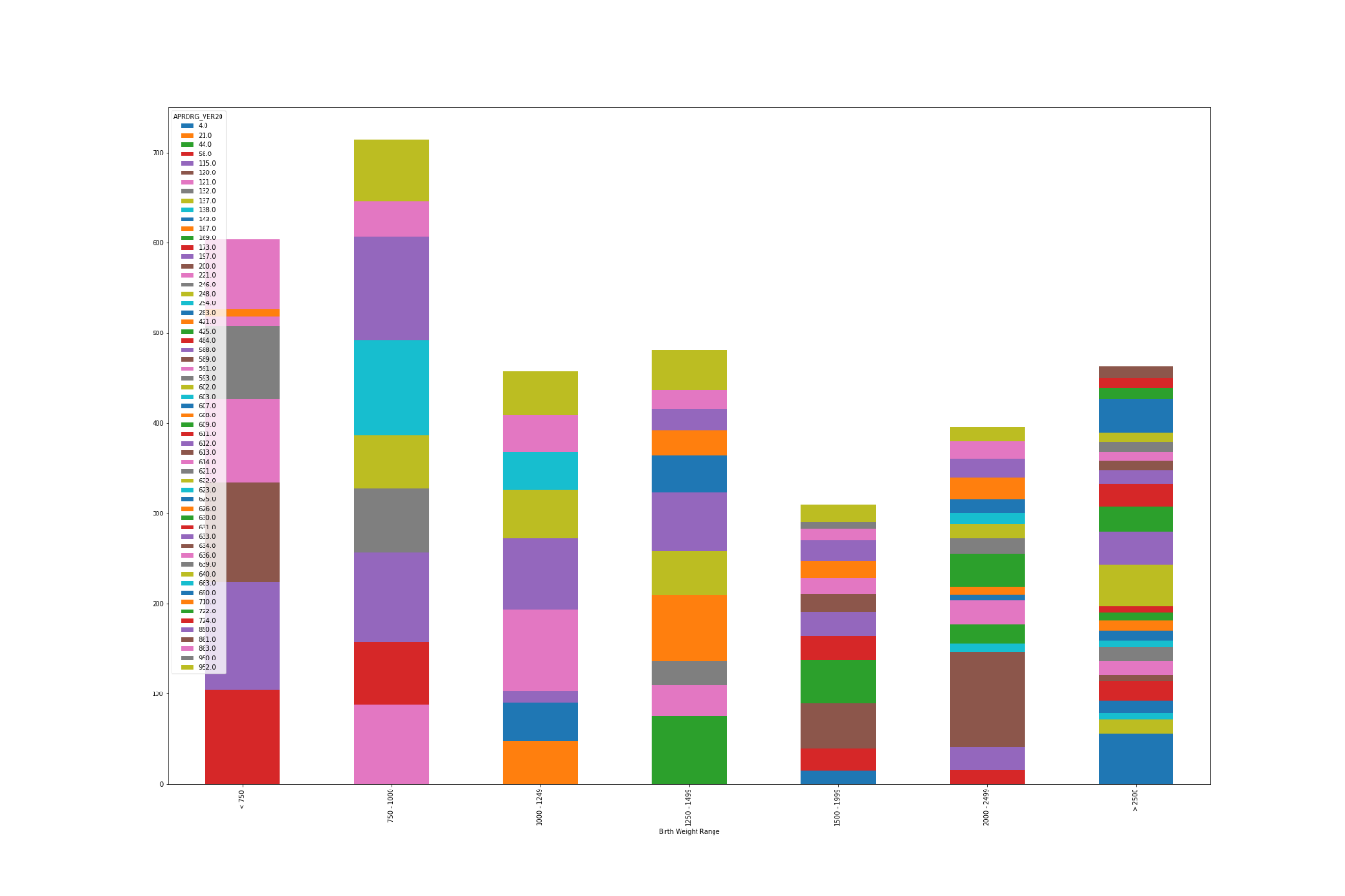
-          1500 - 1999 grams

-          2000 - 2499 grams

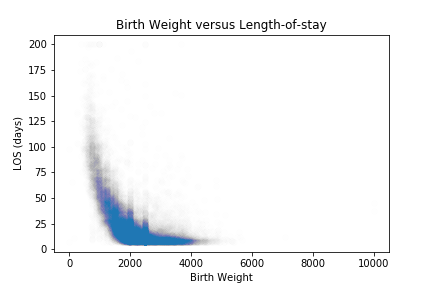
-          > 2500 grams

The table to show the average length of stay vs Birthweight Ranges and APRDRG.

|  |  |  |
| --- | --- | --- |
| **Birth Weight Range** | **APRDRG\_VER20** | **Total mean Los** |
| **< 750** | **173** | **105** |
| **588** | **118.6506849** |
| **589** | **110.0909091** |
| **591** | **92.36027714** |
| **593** | **82** |
| **614** | **11** |
| **626** | **7** |
| **863** | **77.66666667** |
| **750 - 1000** | **121** | **88** |
| **173** | **70** |
| **588** | **98.39072848** |
| **593** | **71.53731343** |
| **602** | **58.23076923** |
| **603** | **67.5** |
| **663** | **38** |
| **850** | **115** |
| **863** | **39.75** |
| **952** | **68** |
| **1000 - 1249** | **21** | **47.5** |
| **143** | **43** |
| **197** | **13** |
| **221** | **90.5** |
| **588** | **78.6746988** |
| **602** | **53.61324786** |
| **603** | **41.32944606** |
| **863** | **42.16666667** |
| **952** | **48** |
| **1250 - 1499** | **44** | **75** |
| **121** | **35** |
| **132** | **26** |
| **167** | **74** |
| **248** | **48** |
| **588** | **65.83050847** |
| **607** | **40.29589041** |
| **608** | **28.62028986** |
| **850** | **23** |
| **863** | **21.33333333** |
| **952** | **44** |
| **1500 - 1999** | **143** | **15** |
| **173** | **24** |
| **200** | **51** |
| **609** | **47.5257732** |
| **611** | **26.73691275** |
| **612** | **26.00057571** |
| **613** | **20.59027778** |
| **614** | **17.5275562** |
| **710** | **19** |
| **850** | **23** |
| **863** | **13.14285714** |
| **950** | **7** |
| **952** | **19** |
| **2000 - 2499** | **58** | **16** |
| **115** | **25** |
| **120** | **105** |
| **138** | **9** |
| **169** | **22.5** |
| **221** | **26** |
| **283** | **7** |
| **421** | **8** |
| **609** | **36.68571429** |
| **621** | **17.64595104** |
| **622** | **15.51701134** |
| **623** | **12.22528736** |
| **625** | **14.88718412** |
| **626** | **11.38702202** |
| **710** | **13** |
| **850** | **21** |
| **863** | **19.33333333** |
| **952** | **16** |
| **> 2500** | **4** | **56** |
| **137** | **15.5** |
| **138** | **7** |
| **143** | **13.66666667** |
| **173** | **22** |
| **200** | **7** |
| **221** | **14.5** |
| **246** | **16** |
| **254** | **8** |
| **283** | **10** |
| **421** | **12** |
| **425** | **8** |
| **484** | **8** |
| **602** | **45** |
| **612** | **37** |
| **630** | **28.39361702** |
| **631** | **24.82585752** |
| **633** | **14.58473625** |
| **634** | **11.10624371** |
| **636** | **9.20959596** |
| **639** | **11.8021164** |
| **640** | **9.507890961** |
| **690** | **37** |
| **722** | **13** |
| **724** | **11.5** |
| **861** | **13** |

The plot to show the average length of stay vs Birthweight Ranges and APRDRG.

Another plot show if there is any relation between Birth weight and Total average length of stay



**Conclusion**

From the above plots, it is observed that bright weight range <750 and 750-1000 have greater average length of stay than other higher birth weight range.

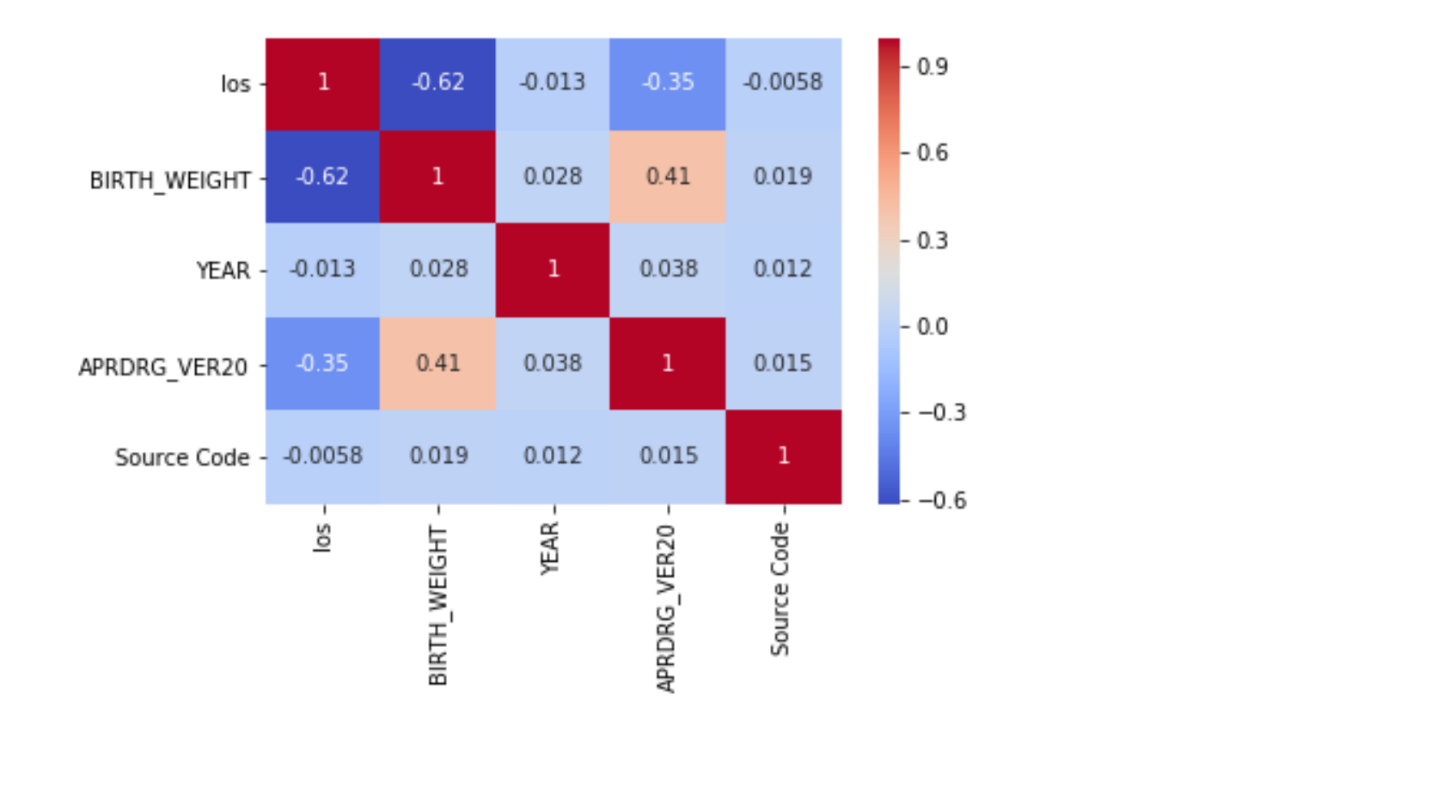
The scatter plot 'Birth Weight versus Length-of-stay' also suggest that birthweight and length of stay are inversely proportion. Hence lower birth weights should have longer length of stays.

**Correlation Heat Maps**

**Data Preparation**

To generate a correlation matrix, all the data fields should have numeral values. Hence a field “Source Code” is generated to map the SOURCE values MANAGED, UNMANAGED and NPIC as 1,2 and 3.

Then a correlation matrix and heat map is created.



**Analysis from Correlation Heatmap**

Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are. That is as one increases so does the other and the closer to 1 the stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonals are all 1/dark red because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and darker the color the higher the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

Here the above correlation matrix indicate some of the relation among variables.

1. The variables Birth\_weight and los forms a square with correlation value as -0.62 which is significant as it is in range of -1 to +1. The (-) indicates that they are inversely proportional to each other. This correlation is also supported by the scatter plot of Birth\_weight vs los

2. The heatmap also shows a significant correlation value of 0.41 between APRDRG\_VER20 and BIRTH\_WEIGHT. That means APRDRG\_VER20 have direct dependency on the Birth Weights.

3. This matrix also reveals that their little correlation (-0.35) between APRDRG\_Ver20 vs los. But since its value is negative, that means these two variables are inversely proportional.

4. There is insignificant correlation between Source Values (Managed-1, Unmanaged-2 and NPIC -3) with any other variables.