Executive Summary

BUDT 758T Group 14

This report is commissioned by the Maryland Department of Health. The hospital's emergency department plays a vital role in healthcare operations and serves as the front line of care for the injured and severely ill. The emergency department is often the only accessible source of providing 24-hour urgent treatment. Thus, it's important to know the potential availability and make better room arrangement to avoid overcrowding.

As a consulting team, we utilize patient information to predict whether a patient will return to a hospital within 30 days of being discharged. With return prediction information, we can prepare in advance to meet further room requirements and make recommendations to alleviate overcrowding. Our project results on patient return prediction will offer operational decision support to the hospital to improve the work efficiency in the emergency department and the quality of healthcare services.

We explore the hospital data using a number of data mining methods. There are 27 variables and 38221 rows in Hospitals_Train data set. We find some problems when exploring variables. First, there are missing values needed to handle, especially large amounts of missing values in some variables. Then, when summarizing variables statistical information, there are very few numbers of data points in a certain level of categorical variables, which may affect the final result because of the unbalanced data distribution. Also, there are many levels in certain categorical variables, which may result in a very complex final model. In addition, based on the definition and statistical result, there are some variables representing similar things or existing multicollinearity. It's also meaningful to create a couple of new variables that can precisely deliver the information of the original variables and reduce the impact of correlation. The solutions will be illustrated in data insight stage.

In modeling insights stage, we utilize a number of the classification analysis methods. We use logistic regression, LDA, and decision tree to find the important variables and make prediction. After setting a cutoff value, we also calculated baseline and compared it with accuracies of models. Based on our project goal, we decide to include TPR as our evaluation method. After exploration, we found that Decision tree had the best performance with 0.7637 accuracy and 0.5721 TPR.

However, we also have some issues that we will have to address as our analysis going forward. We haven't found a proper way to make variable selection. We don't have a strategy to include the relationship between variables into modeling building. We could not decide how to use the time related columns.

Our further work will be:

1. Further exploration of data set and take the relationship of variables into consideration when building models;
2. After variables' relationship exploration, we want to find a way to take advantage of those relationships, such as the way to use RISK and SEVERITY have strong relationship;
3. Figure out how to use the six time-related columns;
4. Explore variable selection methods, and choose variables for each model in validation set;
5. Normalize data when building KNN and perform regularization to prevent overfitting problem when building random forest model;
6. Try new models such as random forest, neural network model and KNN model;
7. Evaluate models and choose the model with best performance based on accuracy (or TPR) in testing data set;

Section 1: Data Insights

We first look through variables one by one to modify data type, handle missing values, regroup or delete multiple levels, remove unnecessary variables and create new variables.

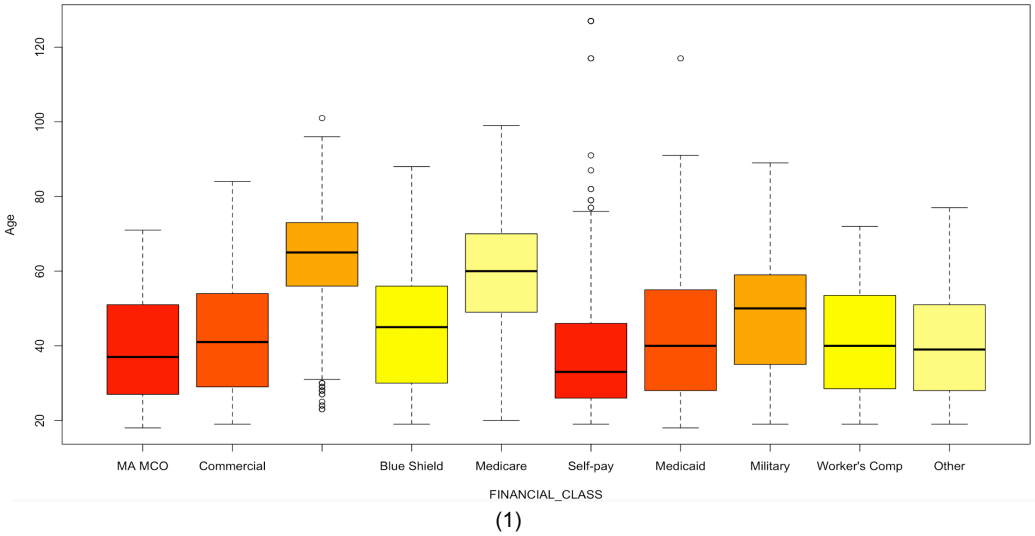
1. *HOSPITAL* is a categorical variable without any missing value.
2. *GENDER* is a categorical variable with two missing values. We remove the missing value directly because it won't affect the result significantly.
3. *AGE* is a numerical variable without any missing value.
4. *RACE* is a categorical variable. There are 177 missing values. We combine missing value and 'Declined to Answer' level together to a new level 'unknown' because whether the missing value or 'Declined to Answer' level, we don't know race anyway. We notice that some levels have a few data points and would not give enough information or would occur error when splitting train, validation and test sets (some level may only have data points in the validation set but not in training set). Thus, we regroup the levels with less than 100 data points to one new level 'Other'.
5. *ETHNICITY* is a categorical variable with 373 missing values. We combine missing value and 'Declined to Answer' level together to a new level 'unknown' because they don't have any clear ethnicity information.
6. *FINANCIAL_CLASS* is a categorical variable without any missing values. We regroup three levels- 'Medicaid Pending', 'Out of State Medicaid', and 'Global Contracts' - into 'Other' because the number of data points is small which might lead to error when we perform models on the different datasets.
7. *WEEKDAY_ARR* and *WEEKDAY_DEP* are both categorical variables with 7 levels and no missing value. We group levels into two new levels: 'Weekday' (Mon - Fri) and 'Weekend' (Sat - Sun).
8. *HOUR_ARR* and *HOUR_DEP* are both categorical variables without any missing value. We group levels into two new levels: 'Day' (8:00 -20:00) and 'Night'.
9. *MONTH_ARR* and *MONTH_DEP* are categorical variables without any missing value. We regroup levels based on season: 'Spring'(3-5), 'Summer'(6-8), 'Autumn'(9-11), 'Winter'(12-2).
10. *SAME_DAY* is a categorical variable without missing value.
11. *ED_RESULT* is a categorical variable with 73 missing values. We delete missing values directly because it won't affect the result significantly. We notice that some levels have a few data points and would not give enough information or would occur error when splitting train, validation and test sets (some level may only have data points in the validation set but not in training set). Thus, we regroup the levels with less than 100 data points to one new level 'Other'.
12. *ACUITY_ARR* is a categorical variable with 3263 missing values. We group missing values to a new level 'unknown'. Also, we delete '5 Purple' level because there is only one data point that can't provide enough information for this specific level.
13. *DC_RESULT* is a categorical variable with 1 missing value.We delete missing value directly because it won't affect the result significantly. There are some levels in this variable containing only a few data points. Like what we do in *ED_RESULT*, we regroup the levels with less than 100 points to a new level 'Others'.
14. *ADMIT_RESULT* is a categorical variable with 30610 missing values. Over 80% of *ADMIT_RESULT* are missing values so it is a huge problem if we want to use it. We set missing values to a new level 'Not Admitted' because if a patient is not admitted, the patient will not have admit result.
15. *CONSULT_ORDER* is a categorical variable without missing values.
16. *CONSULT_CHARGE* is a categorical variable without missing values. However, we notice that there are 37997 data points is '0' and only 224 data points is '1'.
17. *CONSULT_IN_ED* is a categorical variable with 37321 missing values. There is only value of '1' in it. This variable measures whether a consult occurred in the ED or not. There is only a record when the consult actually originated in the ED, so "missing" is the same as "0" in this case.

18. *DIAGNOSIS* is a categorical variable without missing value.
19. *DIAG_DETAILS* is a numerical variable without missing value.
20. *RISK* is a categorical variable with 33045 missing values. Over 80% of *RISK* are missing values so we decide to set them to a new level 'unknown'.
21. *SEVERITY* is a categorical variable with 33045 missing values. Over 80% of *SEVERITY* are missing values so we decide to set them to a new level 'unknown'.
22. *CHARGES* is a numerical variable with 100 missing values. We delete missing value directly because it won't affect the result significantly.
23. *RETURN* is a categorical variable with 141 missing values. We delete missing value directly because it won't affect the result significantly.

- After discussing the relationship between *CONSULT_ORDER* and *CONSULT_CHARGE*, we find that some patients who have consultations by a specialty physician might be charged or might not be charged. We decided to create a new column which was named as *CONSULT_ED* based on the relationship between *CONSULT_ORDER* and *CONSULT_IN_ED*. This new variable means that was a consultation by a specialty physician requested for this patient and did this consultation originate in the ED (0 = no consultation, 1 = has a consultation but not originate in ED, 2 = has a consultation and originate in ED). After doing this, we decide to delete *CONSULT_IN_ED* variable.
- Then, we know missing values in *CONSULT_IN_ED* are the same as "0" in this case. After filling all missing value with 0s, we find that some patients who have consultations by a specialty physician might get the consultation from ED or might not from ED. After analyzing the relationship between *CONSULT_ORDER* and *CONSULT_CHARGE*, we create a new variable named *CONSULT_C*. This variable means that whether a consultation by a specialty physician was requested for this patient and whether the consultation was charged (0 = the patient has no consultation at all, 1 = the patient has consult but not charged, 2 = the patient has a consultation and has been charged).

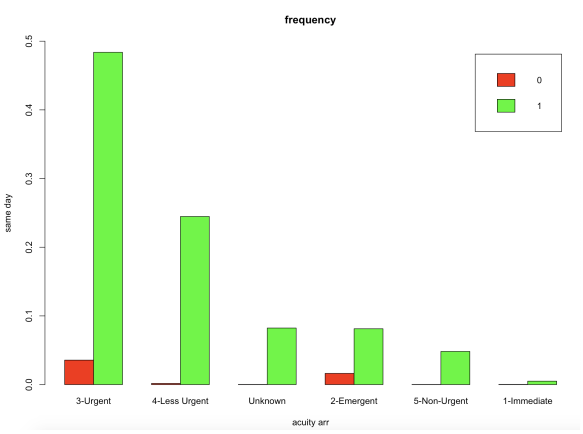
Then, we discover and visualize the relationships among predictors and check potential multicollinearity with the combination of domain knowledge.

We list some interesting relationships we have found as follows:

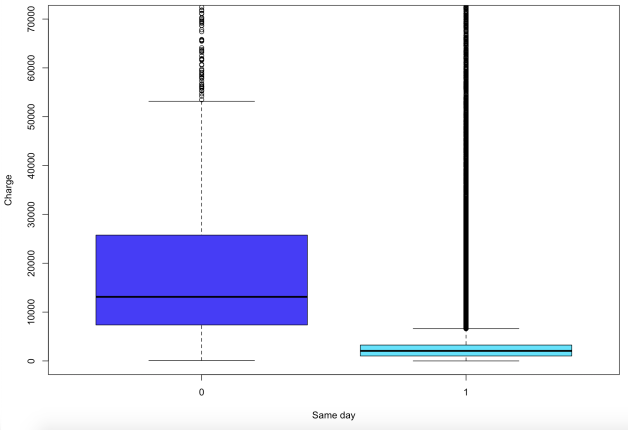


(1)

1.*AGE* and *FINANCIAL_CLASS*:
From plot (1), we can see that older people tend to have insurance such as Replacement Plan and Medicare to cover their health care expenditure, while younger patients may have no insurance to cover and choose self-pay.



(2)



(3)

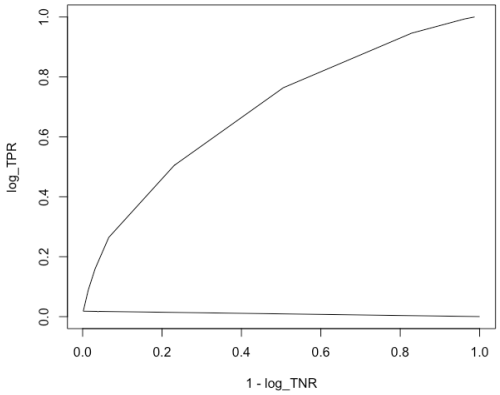
2.*SAME_DAY* and *ACUITY_ARR*:
Considering whether the patient can leave emergency department in one day and his acuity measurement at hospital entry, we can see the influence of patients' acuity to the usage of the emergency department. This can be used to assess emergency department utilization. From plot (2), When *ACUITY_ARR* is '2-Emergent' and '3-Urgent', the percentage that patients leave in different day is relatively high. Emergent and Urgent patients may be the majority that occupies ED room.
3. *SAME_DAY* and *CHARGES*:
From plot (3), the more days the patients stay in hospital, the more charges they will have.
4. *RISK* and *SEVERITY* have very similar relationships with other variables, such as *ACUITY_ARR, CHARGES* and etc.
5. *MONTH_ARR* and *MONTH_DEP*; *WEEKDAY_ARR* and *WEEKDAY_DEP*; *HOUR_ARR* and *HOUR_DEP*, the values in three pairs are almost the same respectively.
6. *CHARGES* has strong relationships with many variables, such as *FINANCIAL_CLASS, ED_RESULT, ACUITY_ARR, DC_RESULT, ADMIT_RESULT*.

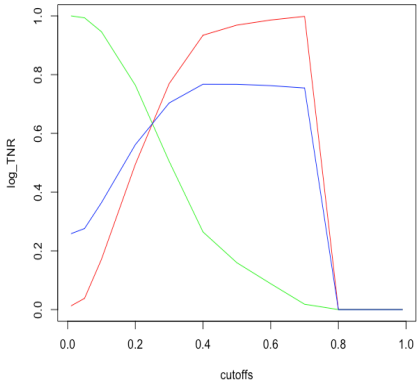
Section 2: Modeling Insights

Before we build any models, we split data set into three sets randomly: 70% of data as training data set, 15% as validation data set and 15% as test data set. Our target variable (*RETURN*) is a categorical variable so our problem is a classification problem. Therefore, models we decided to use is that logistics regression, LDA, Decision tree, and Random Forest. In this phase, we choose to use all predictors which are processed in the previous part.

We first try logistic regression model because it's a common used classification model and we want to quickly set up end-to-end pipeline of the prediction. We set the cutoffs as 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99. After trying different cutoffs and different evaluation methods, we get the 'spider plot' and ROC of the logistics regression model.

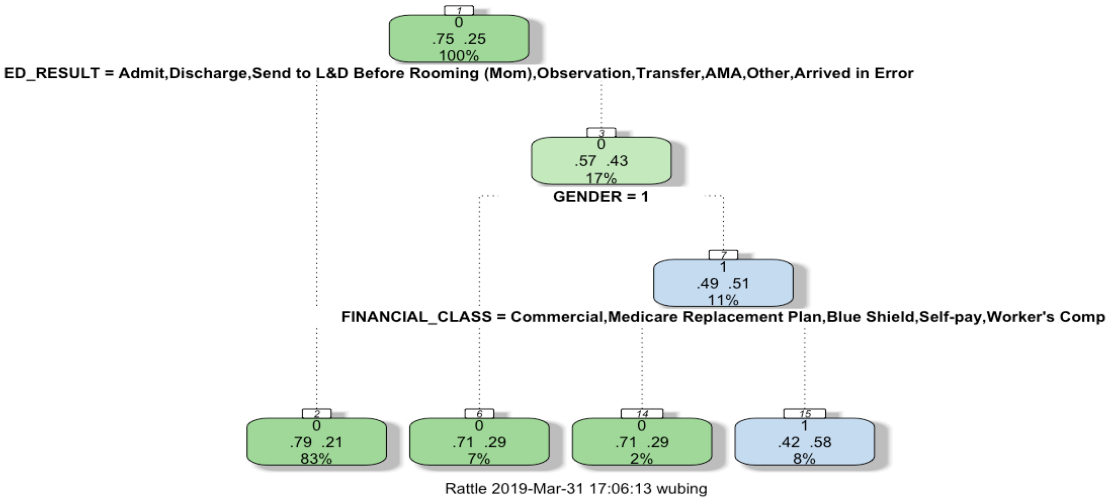
If we use accuracy as the evaluation method, logistics regression model has the highest accuracy at cutoff 0.4. Therefore, we decided to try other models using cutoff 0.4 as well. However, we believe that True Positive Rate (TPR) is another good way to evaluate and compare our models because when we make the prediction for the patients who will come back within 30 days, we want to make sure that our model has the ability to capture as many return patients as possible so that hospital could prepare enough beds for potential patients. Thus, both of them will be considered in our project.





Next, we decide to try two Discriminant Analysis methods which are LDA and QDA because they are data-driven methods for classification problem. However, when we try to run a QDA model, an error occurred in R saying that 'rank deficiency in group 0'. We discuss and conclude that it may be caused by too little data in several group levels or there may be data replication in dataset, and there is insufficient information contained in our data to estimate QDA model.

Then, we decide to try decision Tree because decision tree finds attributes that give information on the target variable. The size of tree is 4. From the tree, we find ED_RESULT is the most important variable since it's the first split of the tree.



Each data set has its own baseline, so we calculate most common baseline in three sets separately. The following list is the information on accuracy of baseline and different models and TPR of different models in different data sets.

Model	Train set accuracy	Train set TPR	Valid set accuracy	Valid set TPR	Testing set accuracy	Testing set TPR
Baseline	0.7526	/	0.7604	/	0.7552	/
Logistic	0.7653	0.2519	0.7674	0.2643	0.7680	0.2611
LDA	0.7650	0.2494	0.7659	0.2660	0.7662	0.2582
Tree	0.7662	0.5762	0.7637	0.5721	0.7691	0.5575

Section 3: Conclusion and Further Work

Conclusion:

After model building, we find the accuracy of models are all slightly above the baseline accuracy which is very disappointing and the TPR values are very low as well. We suspect that our data preprocessing stage hasn't been done properly and we might not tune models using the right way. However, we are happy to find some interesting stories behind the data set and understand the operation of healthcare industry, which is a great opportunity for us to understand healthcare information and preparing us for the future job market. Moreover, the data exploration progress is very challenging because of how dirty the data is that make this practice more realistic.

According to the results from logistic regression, our conclusions are listed below:

1. LWBS individuals are more likely to return
2. Individuals with severe ACUITY_ARR are less likely to return
3. The more the individual is charged, the less likely he/she will return
4. Hospitals don't influence much on the patient's return.
5. Hospitals need to upgrade their time recording system.

Recommendation:

1. Hospitals and insurance companies can have tiered pricing system. For individuals who have minor risk and severity, we can charge them more if they enter ER.
2. Hospitals can adjust their schedule and budget of ER based on the model.

Further Work:

In phase II, the first main work is variable selection.

First, we need to consider multicollinearity and correlation for models we have built and take the relationship of variables into consideration. After variables' relationship exploration, we want to find a way to take advantage of those relationship. For example, according to what we analyzed in data insight section and relationships among predictors, we need to add interaction variable of *SAME_DAY*CHARGES* to models. Besides, considering that *RISK* and *SEVERITY* have very similar relationships with *ACUITY_ARR and RISK* and *SEVERITY* have too many missing values, we may remove *RISK* and *SEVERITY,* and only include *ACUITY_ARR* in our models.

Second, we need to consider normalization and regularization to improve model performance. For example, the variation of *CHARGES* is very large, we want to use regularization method to process data, such as scaling, taking the logarithm, and etc.

Then, we have to figure out how to use the six time-related columns. *MONTH_ARR* and *MONTH_DEP*; *WEEKDAY_ARR* and *WEEKDAY_DEP*; *HOUR_ARR* and *HOUR_DEP*, the values in three pairs are almost the same respectively. We already have variable *SAME_DAY* as a measure of the time patients stayed in emergency room, we will delete *MONTH_DEP, WEEKDAY_DEP,* and *HOUR_ARR.*

Finally, after variable selection, we will choose variables for each model in validation set.

The second main work is building more models. First, we will try random forest, which is not only a model for prediction, but also can help us in further variable selection. Second, we will also include KNN, before this we need to change all the categorical variables to dummy variables, then find the best k-value for further

comparison. Finally, neural network is another model we want to use, we will use deep learning and backpropagation method to predict the possibility a patient will return or not.

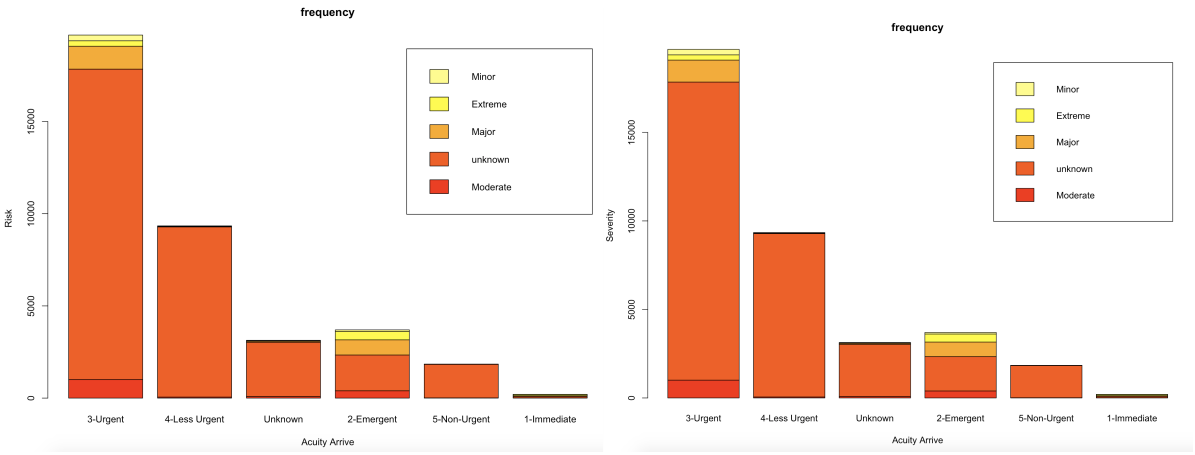
The last main work is model selection and prediction. We will use testing data set to choose the model with best performance for prediction. Then, we will predict new, unseen data in testing data file.

Group Member Roles and Responsibilities

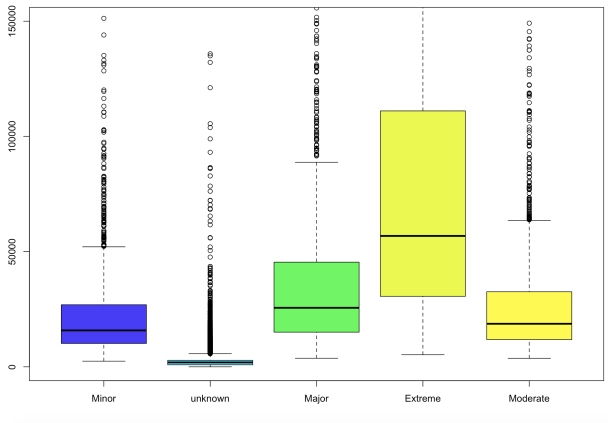
- Cleaning data (pg. 2,3): Zheming Chen, Yefei Fan
- Creating new variables and variables removing (pg. 3): Bing Wu
- Visualizing data relationships, (pg. 3,4): Yingjie Gao
- Logistics regression (pg. 4,5): Zheming Chen, Yingjie Gao
- LDA & QDA(pg. 5): Yefei Fan
- Running and evaluating, decision tree model (pg. 5): Bing Wu

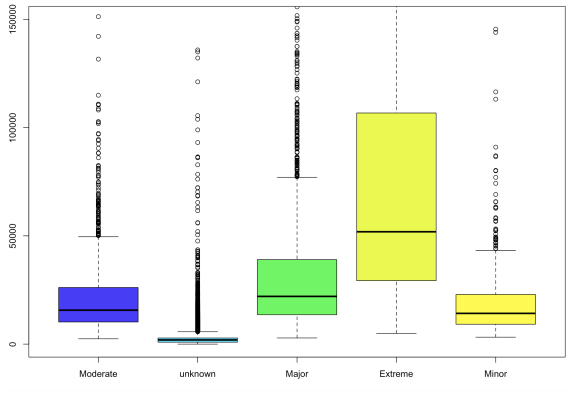
Appendix

- Relationship between *RISK* and *ACUITY_ARR* vs. Relationship between *SEVERITY* and *ACUITY_ARR*

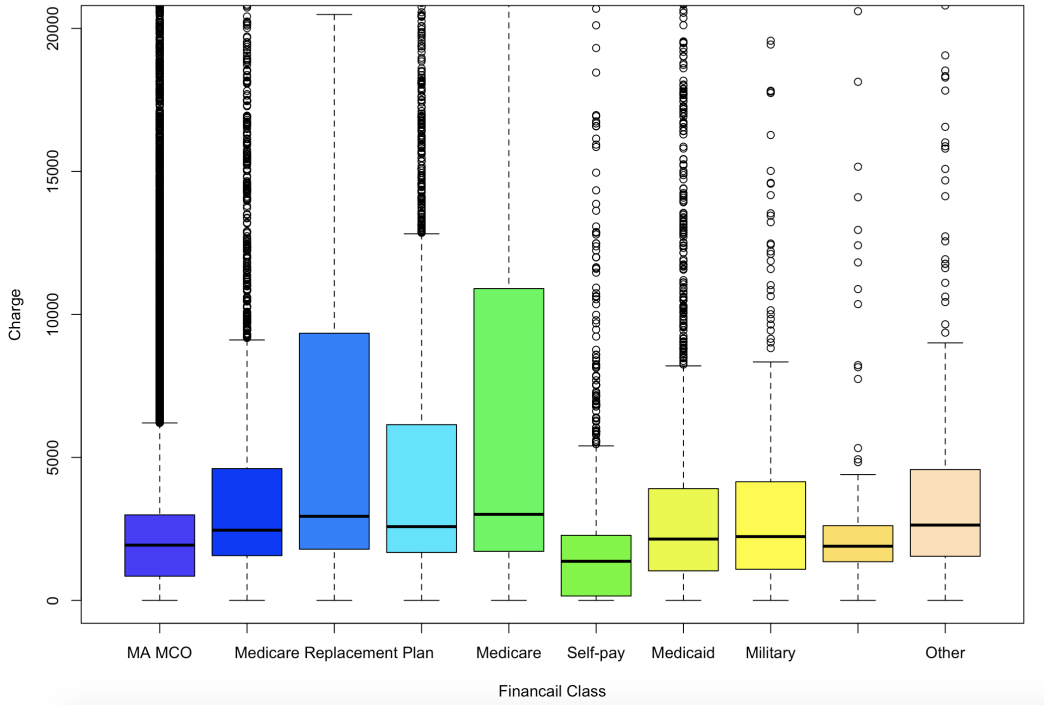


- Relationship between *RISK* and *CHARGES* vs. Relationship between *SEVERITY* and *CHARGES*

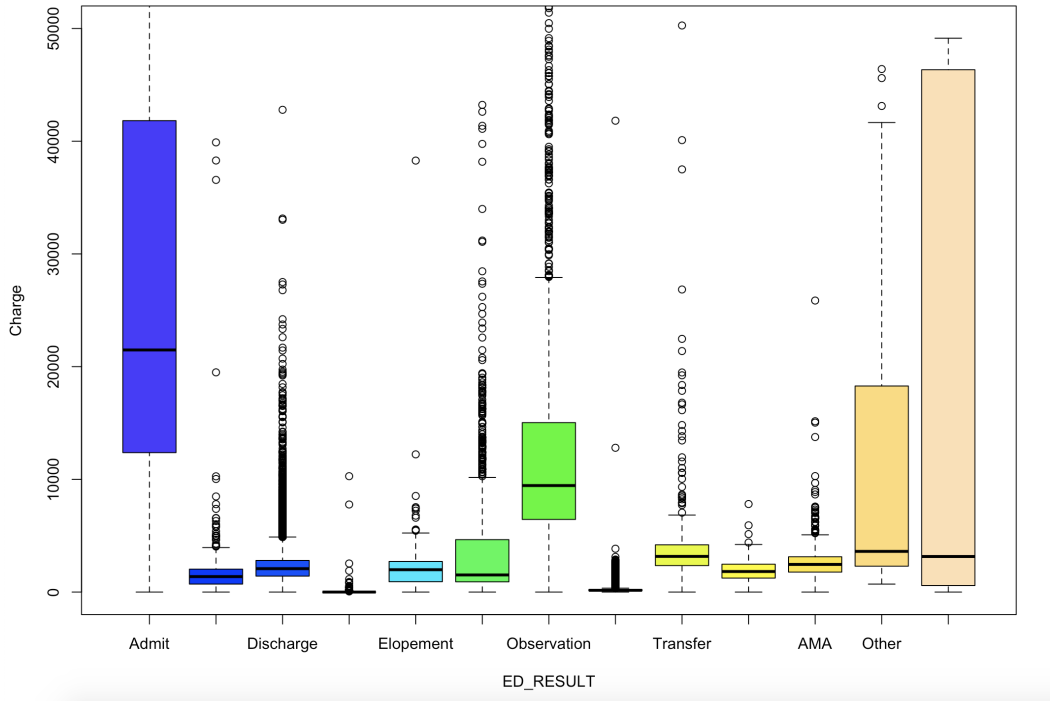




- Relationship between *CHARGES* and *FINANCIAL_CLASS*



- Relationship between *CHARGES* and *ED_RESULT*



- Relationship between *CHARGES* and *ACUITY_ARR*