Shuhua (Jessica) Yin Project Portfolio

MS in Data Science and Business Analytics, UNC Charlotte
BS in Statistics, NC State University

Project 1: Graduate Research Project: Cardiotoxicity
Project 2: Teradata Challenge 2018: Bike MS
Project 3: Analytics Modeling on Claim Datasets

Project 1 (with example visuals) January 2018 – Present

Objectives & Background:

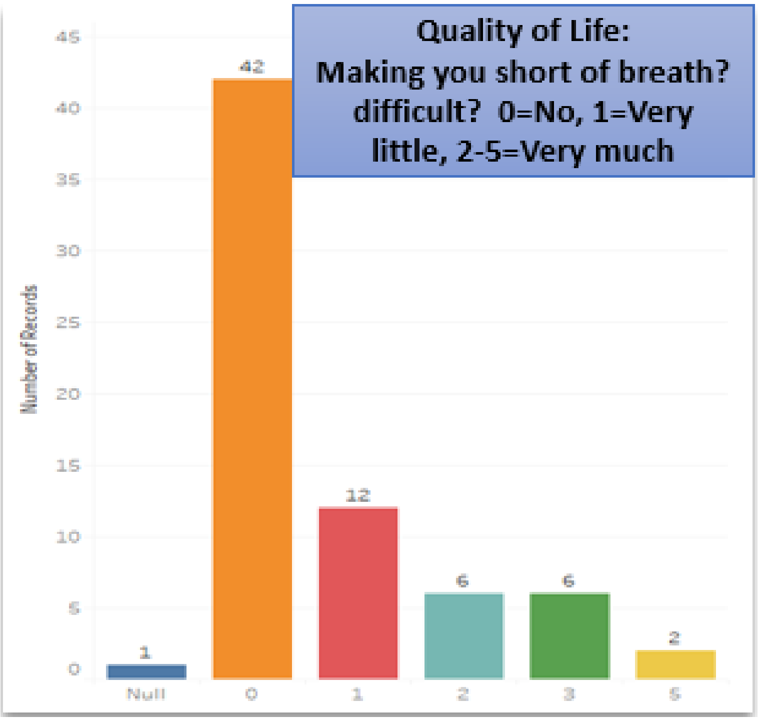
- Anthracycline-based chemotherapy aims to erase the undetectable cancer cells and reduce recurrence, but caused cardiovascular abnormalities
- Data: 224 breast cancer patients with information and measurements
- Aims to predict if and how likely a cancer patient is going to have heart issues in long term (on month 24) based on baseline (month 0) information

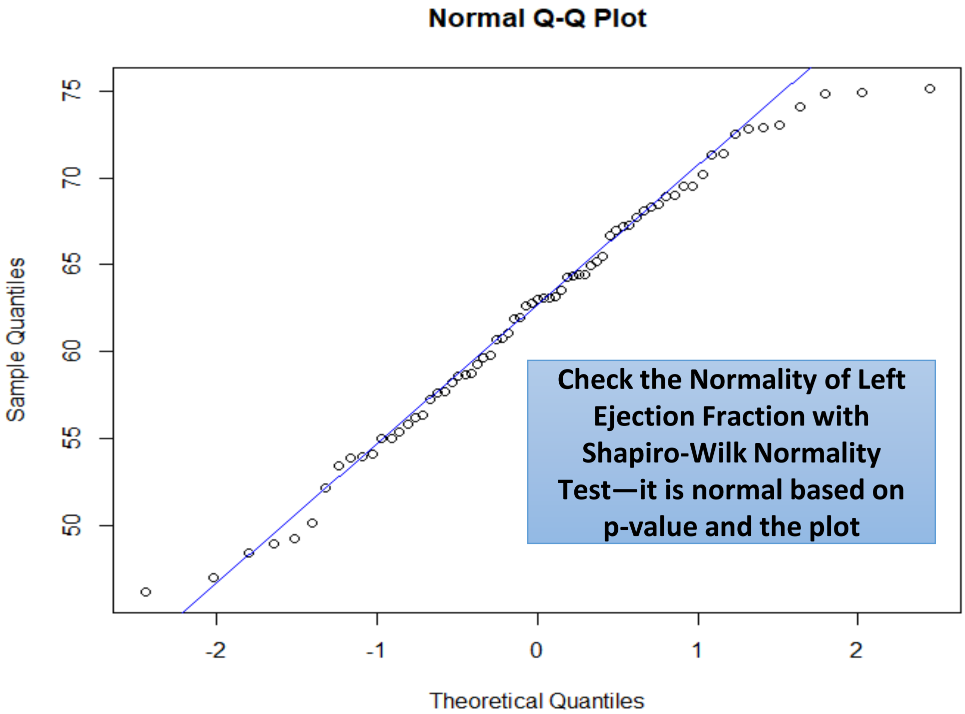
Data Wrangling & Exploration (R, Python, Tableau):

- Filter out columns with desired percentage of missing values
- Slice and dice data from different perspectives and use the one that provides the most info
- Explore distributions, normality tests, correlation tests on all variables

Data Analysis & Predictive Modeling (on-going; R, Python):

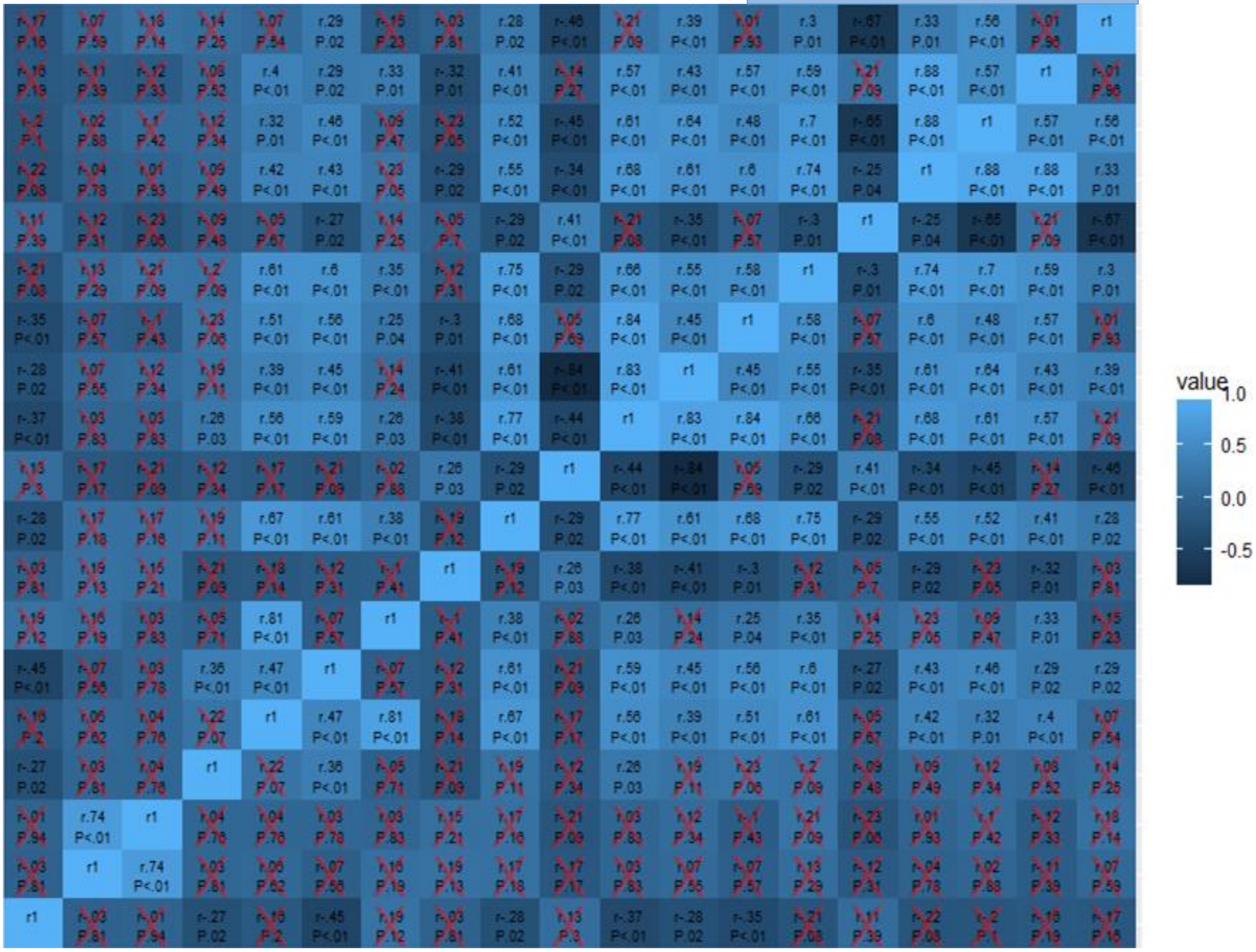
- Build different machine learning models (Logistic Regression, Linear Regression, LASSO, Ridge Regression, Elastic Net, Neural Network, Random Forest)
- Evaluate the accuracy of each model on our data and select the most fitted method
- Fit the model with our data and produce the desired predictions





Check the Normality of Left Ejection Fraction with Shapiro-Wilk Normality Test—it is normal based on p-value and the plot

Correlations and their significance among all the variables



Project 2 (with example visuals) February 2018 – May 2018

Objectives & Background:

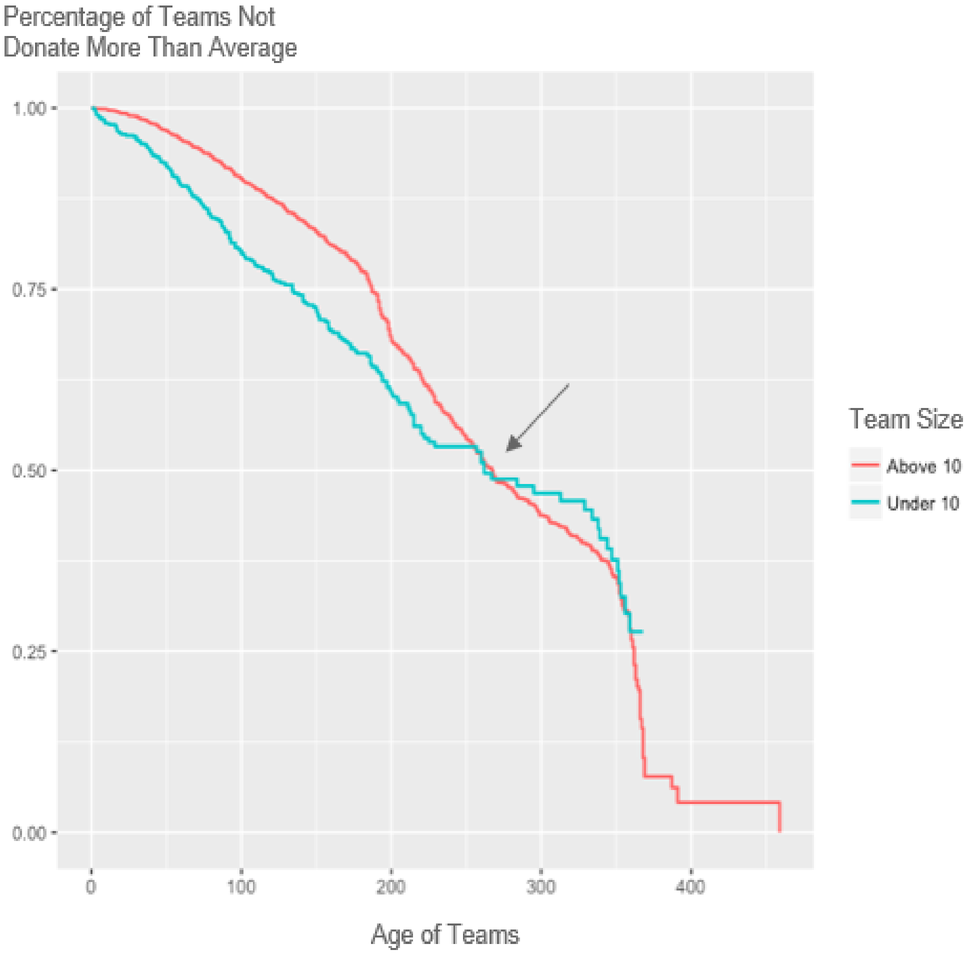
- Fundraising campaign Bike MS has a steady decline since 2012 due to lack of participants
- 2018 challenge hosted by Teradata aims to find features that significantly impact on the donation amount of Bike MS events

Data Wrangling & Exploration (Tableau):

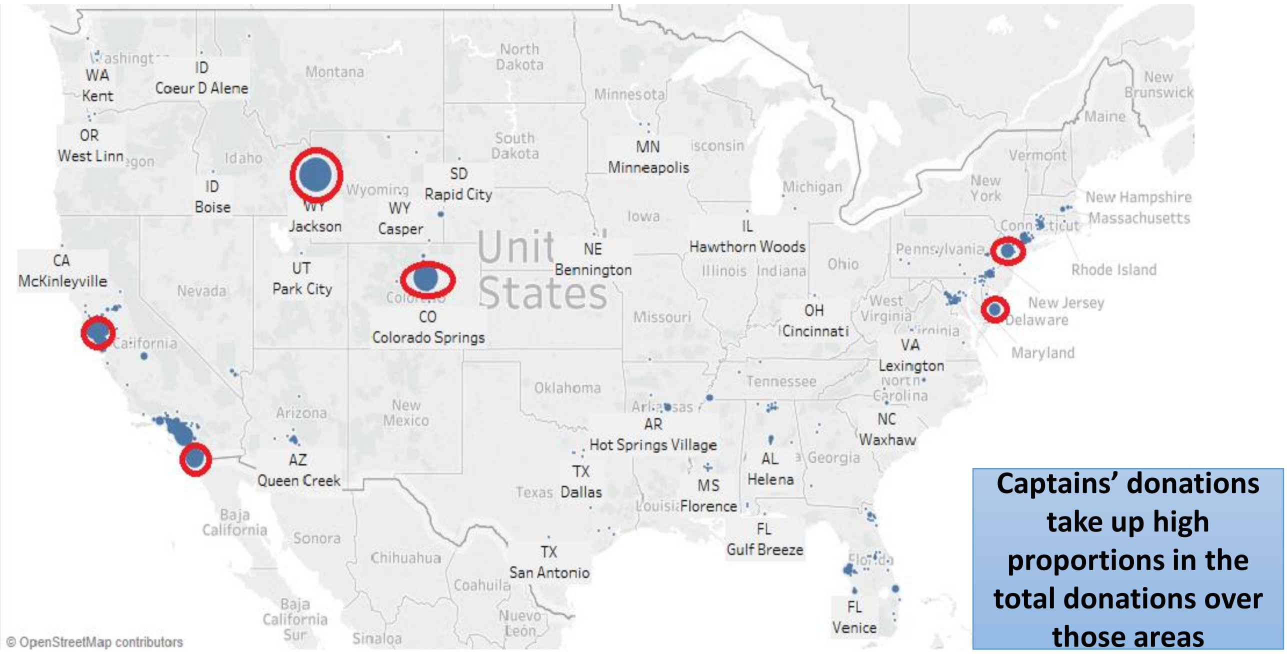
- Select the appropriate datasets from all the given datasets based on our specific goals
- Variable distributions

Data Analysis & Conclusions (R, Tableau):

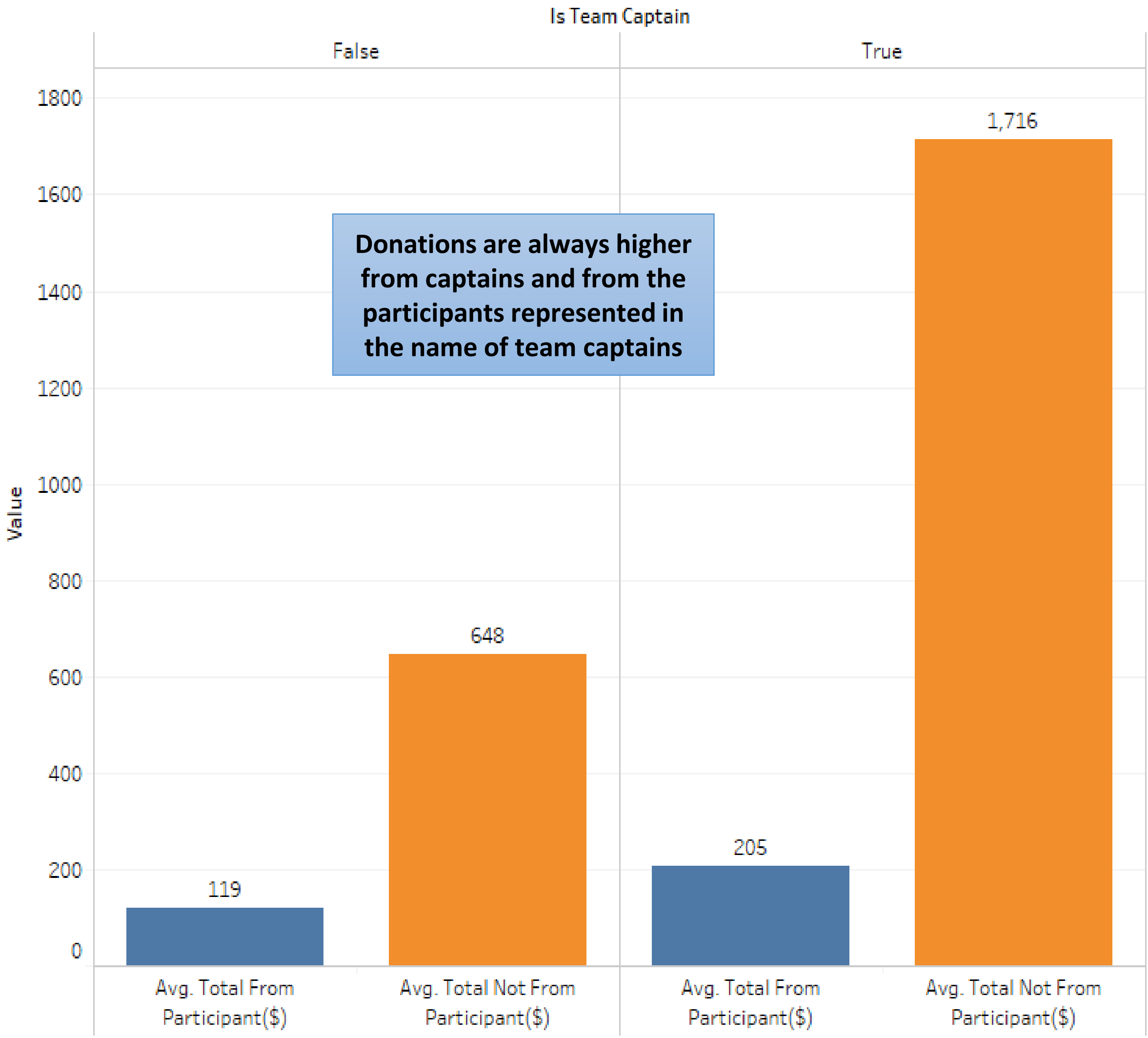
- Use machine learning methods (Neural Network, Boruta, Random Forest, Linear Regression) for feature selections and evaluate all model performances
- According to the significant features, explore their relationships with the donation amount
- Team size impacts the donation amount: team should have less than 10 members in the first 9 month, more than 10 members after the 9th month
- Should target the states and corporates that usually have higher donation amounts
- It is crucial to have captains in the teams, since all teams that had captains had the highest donation amount



Survival analysis shows the importance of team size in different situations



Captains' donations take up high proportions in the total donations over those areas



Donations are always higher from captains and from the participants represented in the name of team captains

Project 3 (with example visuals) October 2017 – December 2017

Objectives & Background:

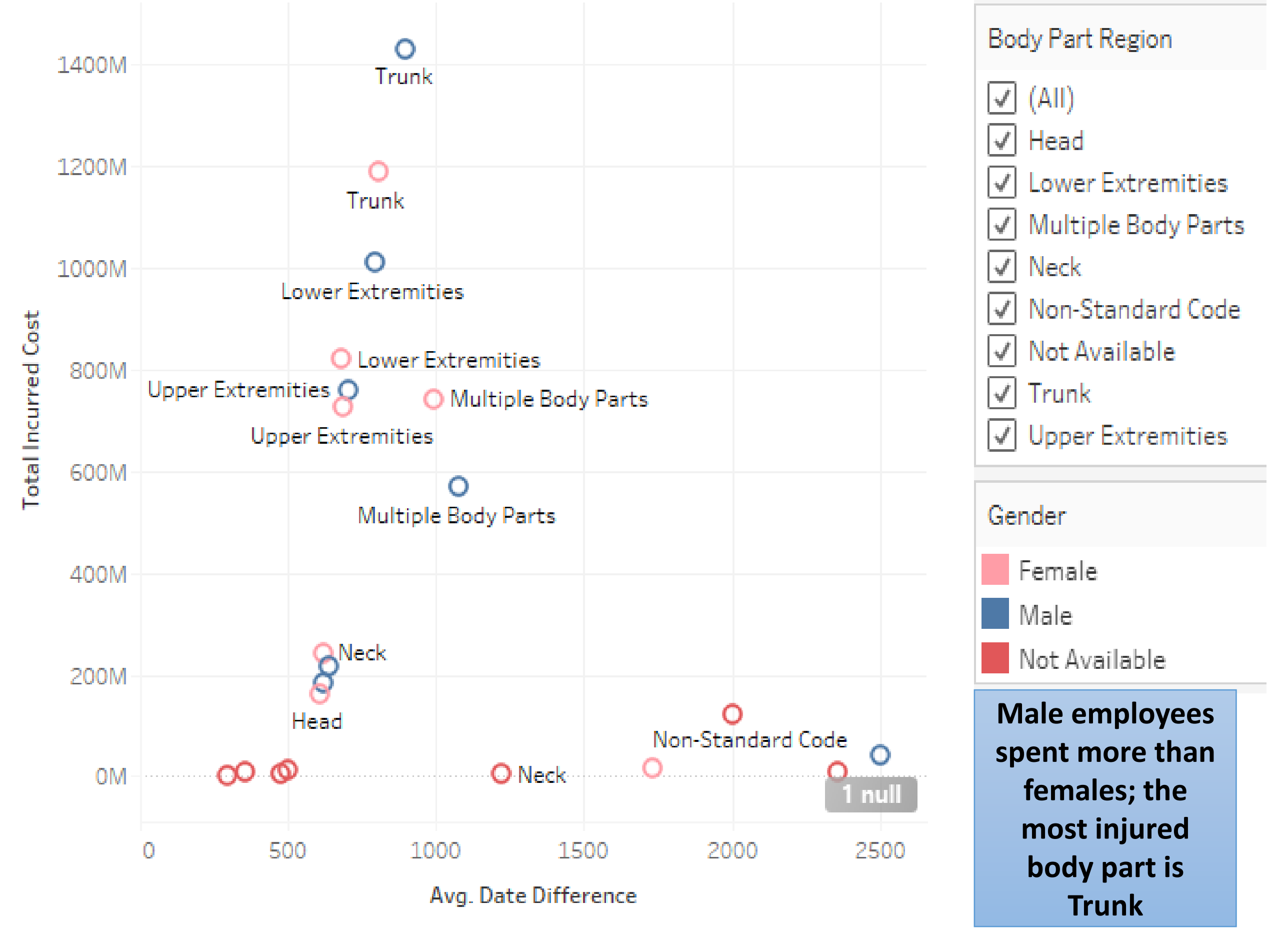
- Worker's compensation claims
- Wants to find the major drives of claim costs & claim process time
- Improve claims management business
- Claim status (open & close), Total Recovery, Total Reserve, Claimant Open Date & Close Date, Injury Nature, Body Part Region, Is Fatal, etc.

Data Wrangling & Exploration (SAS Enterprise Guide, Tableau):

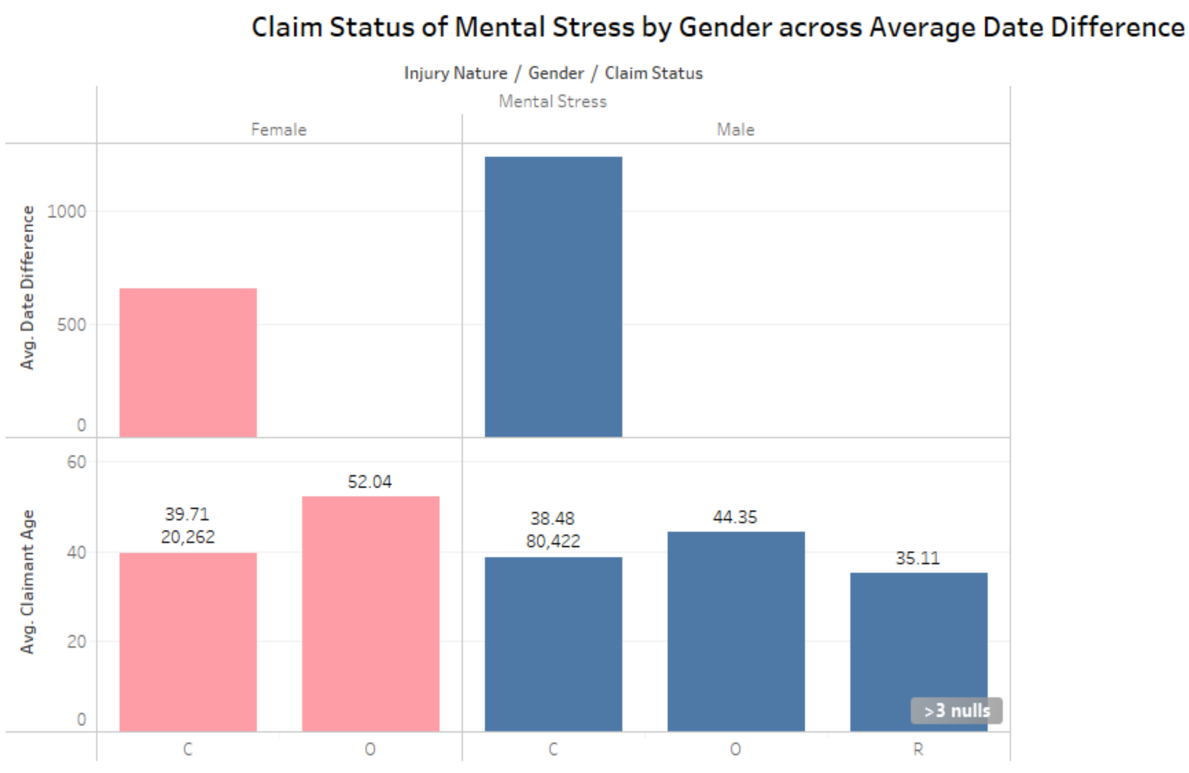
- Combine two given datasets into one only selecting the desired variables
- Roughly explored some relationships in between variables and the claim costs and processing time

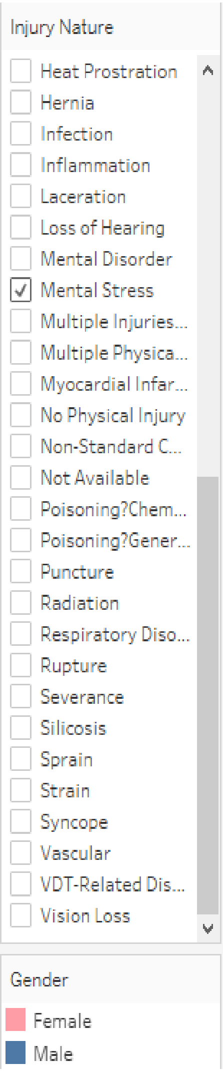
Data Analysis & Conclusions (R, SAS Enterprise Guide):

- Use Logistic Regression to fully examined how variables impact the amount of claim costs and length of claim processing time
- Expedite claims predicted critical for claims processing time
- Recovery time helps strengthen predictive capability of models
- It is essential to expand types of data to be collected and analyzed so more model techniques can be adequate



Male employees spent more than females; the most injured body part is Trunk





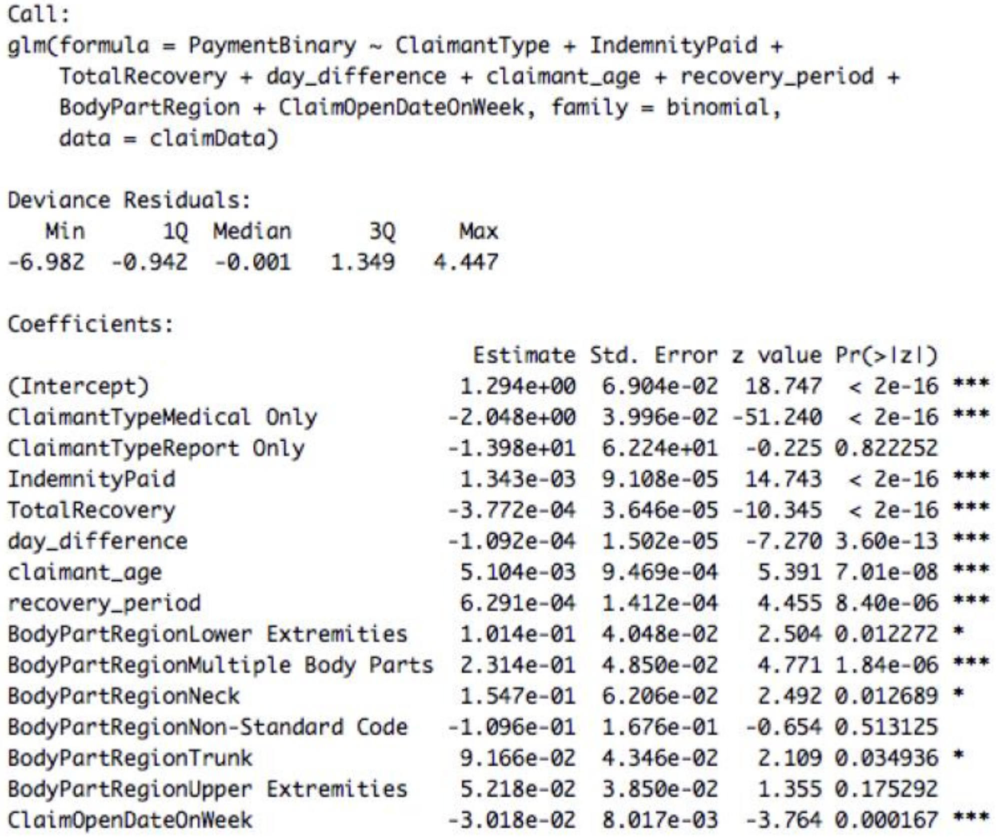
Injury Nature

☐ Heat Prostration
☐ Hernia
☐ Infection
☐ Inflammation
☐ Laceration
☐ Loss of Hearing
☐ Mental Disorder
☑ Mental Stress
☐ Multiple Injuries...
☐ Multiple Physica...
☐ Myocardial Infar...
☐ No Physical Injury
☐ Non-Standard C...
☐ Not Available
☐ Poisoning?Chem...
☐ Poisoning?Gener...
☐ Puncture
☐ Radiation
☐ Respiratory Diso...
☐ Rupture
☐ Severance
☐ Silicosis
☐ Sprain
☐ Strain
☐ Syncope
☐ Vascular
☐ VDT-Related Dis...
☐ Vision Loss

Gender

🟥 Female
🟦 Male

In this case male employees took way longer for their injury claims to close than females on Mental Stress; older people have more open cases on Mental Stress



```
Call:
glm(formula = PaymentBinary ~ ClaimantType + IndemnityPaid +
    TotalRecovery + day_difference + claimant_age + recovery_period +
    BodyPartRegion + ClaimOpenDateOnWeek, family = binomial,
    data = claimData)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-6.982  -0.942  -0.001    1.349    4.447

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.294e+00  6.904e-02  18.747  < 2e-16 ***
ClaimantTypeMedical Only         -2.048e+00  3.996e-02 -51.240  < 2e-16 ***
ClaimantTypeReport Only          -1.398e+01  6.224e+01  -0.225 0.822252
IndemnityPaid                     1.343e-03  9.108e-05  14.743  < 2e-16 ***
TotalRecovery                    -3.772e-04  3.646e-05 -10.345  < 2e-16 ***
day_difference                   -1.092e-04  1.502e-05  -7.270 3.60e-13 ***
claimant_age                      5.104e-03  9.469e-04   5.391 7.01e-08 ***
recovery_period                   6.291e-04  1.412e-04   4.455 8.40e-06 ***
BodyPartRegionLower Extremities   1.014e-01  4.048e-02   2.504 0.012272 *
BodyPartRegionMultiple Body Parts 2.314e-01  4.850e-02   4.771 1.84e-06 ***
BodyPartRegionNeck                1.547e-01  6.206e-02   2.492 0.012689 *
BodyPartRegionNon-Standard Code  -1.096e-01  1.676e-01  -0.654 0.513125
BodyPartRegionTrunk               9.166e-02  4.346e-02   2.109 0.034936 *
BodyPartRegionUpper Extremities   5.218e-02  3.850e-02   1.355 0.175292
ClaimOpenDateOnWeek              -3.018e-02  8.017e-03  -3.764 0.000167 ***
```

Example R output with claim costs being outcome variables; Multiple Body Parts injuries seem to increase more costs; Recovery Period is crucial