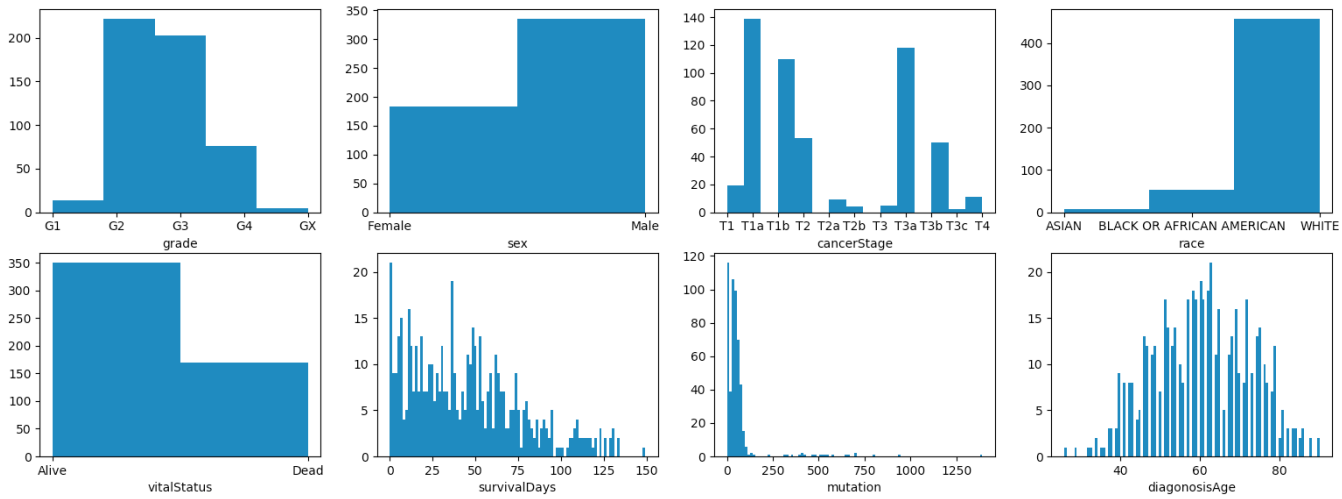
Convergent Genomics Data Science Challenge

Exploratory Data Analysis (EDA)

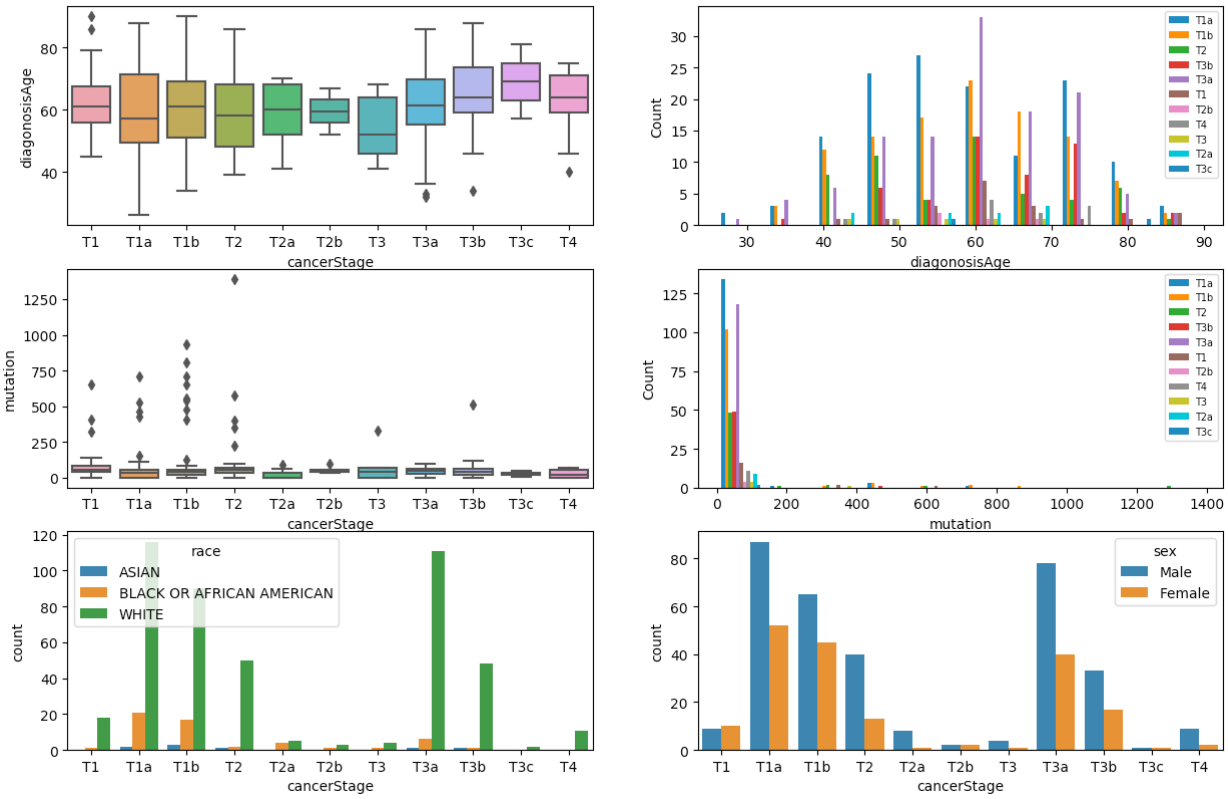
For patient_data.tsv

- Statistics distributions of variables in patients.
    - The target categories of cancerStage, grade and vitalStutus are all un-balanced dataset.
    - There are outliers in mutation columns.

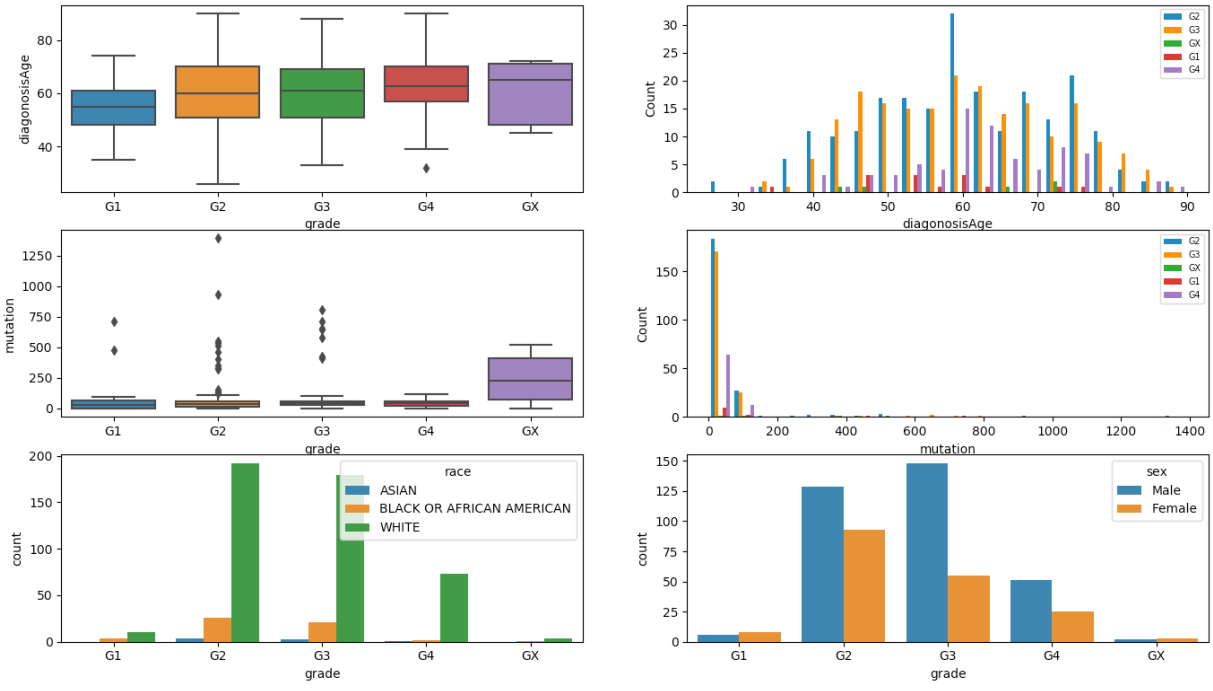


	diagonosisAge	survivalDays	mutation
count	520.000000	520.000000	520.000000
mean	60.521154	44.526173	61.432692
std	12.209457	32.470140	119.998681
min	26.000000	0.000000	0.000000
25%	51.000000	17.912500	19.000000
50%	60.500000	39.420000	41.000000
75%	70.000000	63.420000	60.000000
max	90.000000	149.050000	1392.000000

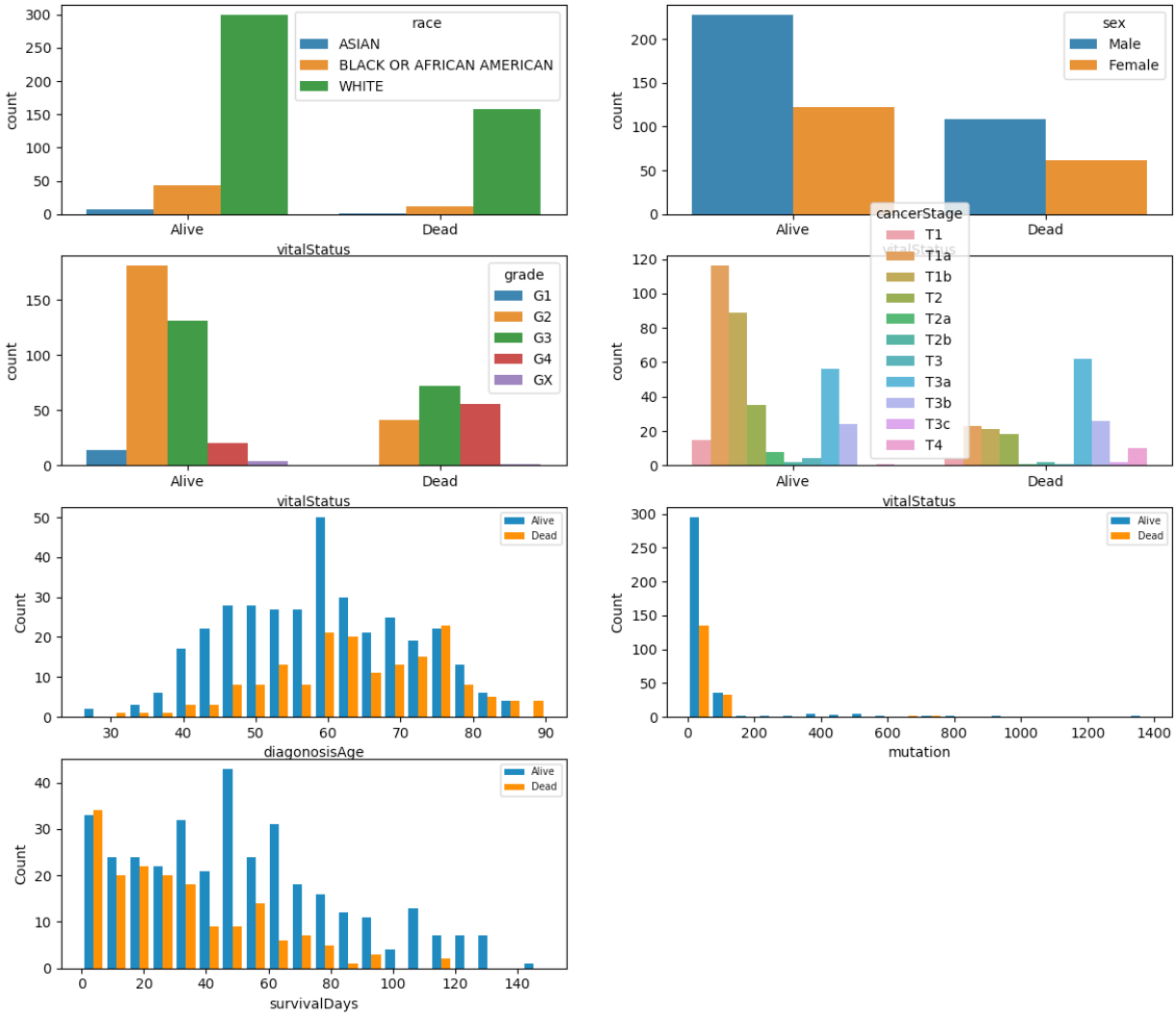
- Data Visualization of Relationship between target variables (cancerStage, grade, vitalStatus, survivalDays) and independent variables from patients
    - cancerStage: It looks diagonosisAge could be a feature of cancerStage.



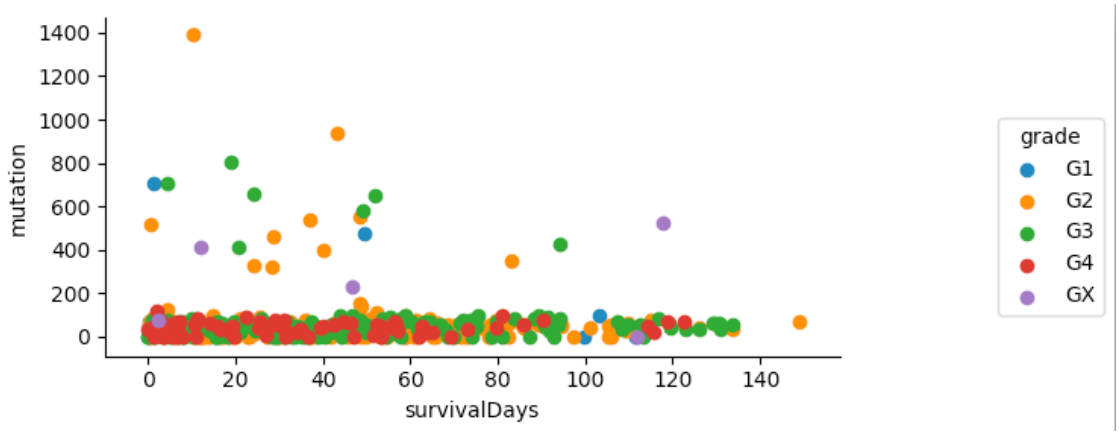
- grade: It looks diagonosisAge and mutation should be a feature to predict grade

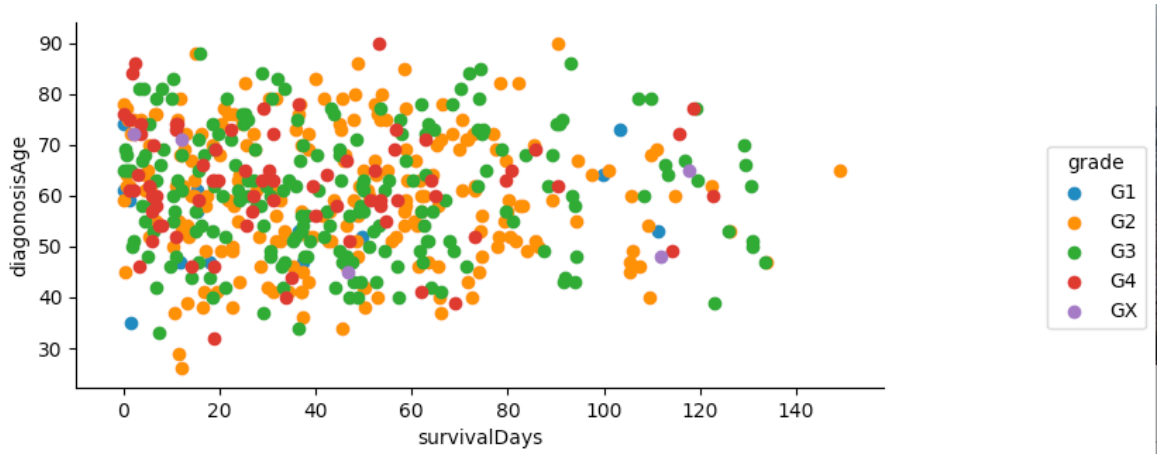


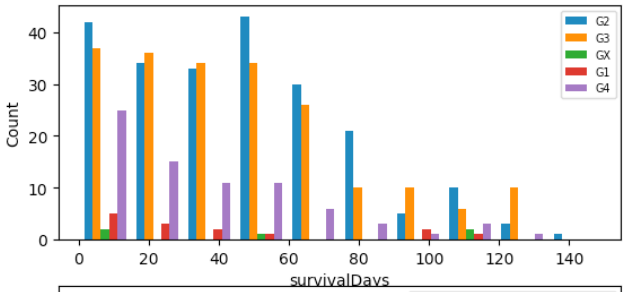
- vitalStatus: grade, cancerStage, survivalDays, diagonosisAge and mutation all look related with vitalStatus

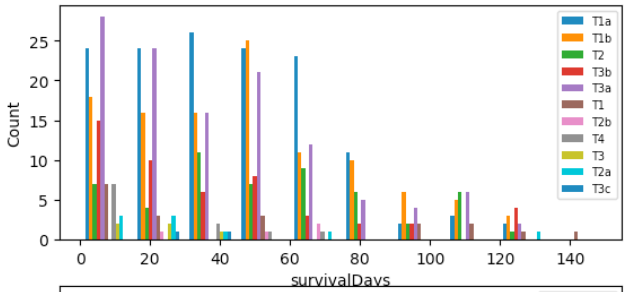


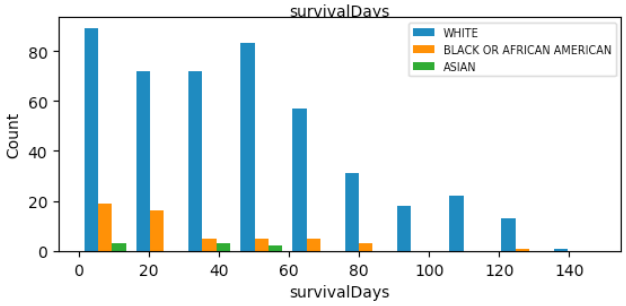
- survivalDays: mutation and diagonosisAge are mixed together, looks they are not good to be selected as features. grade and cancerStage should be selected from figure and common sense.

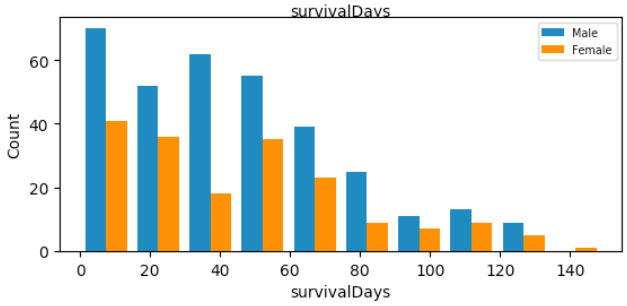

mutation
survivalDays
grade
G1
G2
G3
G4
GX


diagonosisAge
survivalDays
grade
G1
G2
G3
G4
GX


Count
survivalDays
G2
G3
GX
G1
G4

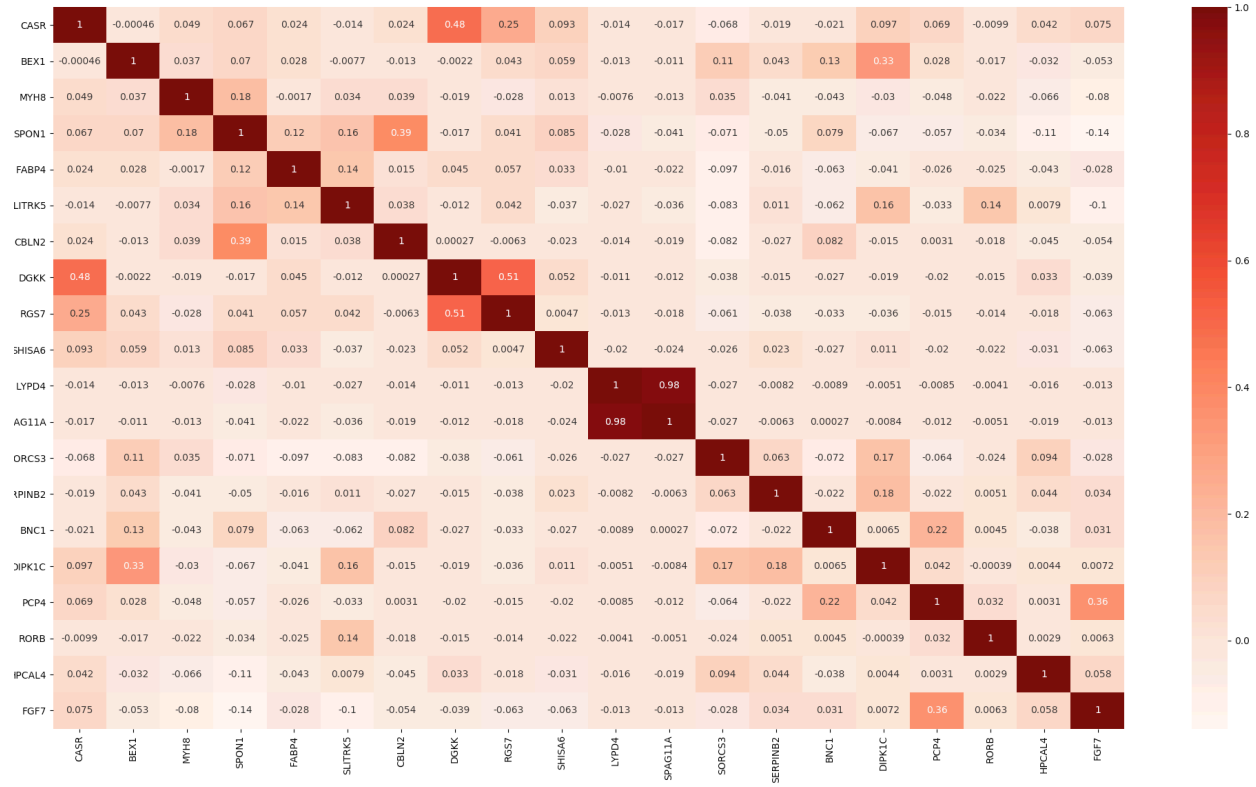

Count
survivalDays
T1a
T1b
T2
T3b
T3a
T1
T2b
T4
T3
T2a
T3c


Count
survivalDays
WHITE
BLACK OR AFRICAN AMERICAN
ASIAN


Count
survivalDays
Male
Female

Correlation analysis for mrna_data.tsv

- Compute pairwise correlation of columns, based on ‘pearson’ standard correlation coefficient. The figure only shows correlation from 20 columns (80 no-null columns in total) for better understanding. The results shows most pairwise correlation is weak, means there are no strong linear relationship between columns. Pairs of (LYPD4, AG11A), (DGKK, RGS7), (CASR, DGKK), (SPON1, CBLN2) have the most strong relationship for 20 columns. PCA for dimension reduction will be discussed after.



For seq_data:
Only a few records have data.

```
       STUDY_ID      SAMPLE_ID   CASR   MYH8        SPON1   SLITRK5  LYPD4  \
count       534            534      5      4            1         8      1
unique        1            534      5      4            1         5      1
top     kirc_CG  CG-B0-5702-01  F788Y  D162Y  X543_splice  V8Sfs*13  F181L
freq        534              1      1      1            1         3      1

       SPAG11A  SORCS3   BNC1      DIPK1C   RORB  FGF7       ROS1 SLC6A17  \
count        1       1      3           2      1     2          9        2
unique       1       1      3           2      1     2          8        2
top       R60W  M1129I  V960E  S328Qfs*15  D348N  T27S  I414Nfs*4   S587I
freq         1       1      1           1      1     1          2        1

         SCG2  KCND2 CAPSL    F12 TCEAL2 SLITRK6 DEFA4 SORCS1 CHRNA4 ABCG8  \
count       5      4     1      3      1       9     1      1      2     3
unique      5      4     1      3      1       9     1      1      2     3
top     A320S  R254H  P25R  Q294H  N143K   A101V  S59P  D508E  R566Q  T76A
freq        1      1     1      1      1       1     1      1      1     1

       SLC17A8  CD5L TRIM63 DPEP1  KCTD8 C16ORF78 PRSS12       MYH4 USH1G  \
count        2     1      1     1      1        2      3          8     1
unique       2     1      1     1      1        2      3          7     1
top      T576I  V37G  S202R  R96W  K127R     D99G  G800S  K59Efs*28  D11A
freq         1     1      1     1      1        1      1          2     1

             RLN1  EPHA7 STARD6       NTRK2  FOXP2  KCNJ1  MS4A3 OR51B4  \
count           1      1       1           1      3      1      2      2
unique          1      1       1           1      3      1      2      2
top     V40Sfs*27  Q809H   P58T  D275Ifs*33  Q131H  F173V  Q119H  V247L
freq            1      1       1           1      1      1      1      1

            GALNT9  KCNK3 PCDHB1 LRRTM3
count            1      1      3      3
unique           1      1      3      3
top     P147Lfs*25  G117D  G428R  N504K
freq             1      1      1      1
```

Data Preprocessing

For patient_data.tsv

- Drop columns with same values: cancerType, histologicType, samples (only one record with 2, all others 1), profiledAlter
- Drop ‘survivalStatus’, since it has the exactly the same value as ‘vitalStatus’
- Convert columns to float type: diagonosisAge, survivalDays, mutation
- Convert columns to category: profiled, informed, race, sex, cancerStage, grade, vitalStatus
- Drop records if grade or vitalStatus is None, since they are prediction target
- Drop records if race is None, since only less than 10 records
- Fill in by ‘No’ for records with None profiled, then it will be treated as one category
- Change MALE as Male for ‘sex’ column

For mrna_data.tsv

- Drop columns with all none values (80 left)
- Drop STUDY_ID with same values

- Cut SAMPLE_ID to string with length 10. For example, from CG-B0-5710-01 to CG-B0-5710, which will be used to merge with patient data.

Feature Transform / Normalization / Scaling

- Apply preprocessing.StandardScaler (z-score) to scale diagonosisAge and survivalDays, and all columns in mrna.
- Apply preprocessing.RobustScaler to mutation with outliers
- Apply preprocessing.LabelEncoder to categorical variable to transform them to integers, then apply OneHotEncoder for feature embedding

Dimension Reduction

- Try with decomposition. PCA to do dimension reduction for mrna data by function *PCA4Mrna(df4Mrna, varianceRatio).*
    - When varianceRatio set to 95, feature decreased from 80 to 63
    - When varianceRatio set to 90, feature decreased from 80 to 54

The feature numbers did not get big decreased since there does not have many strong linear relationship between features in mrna.

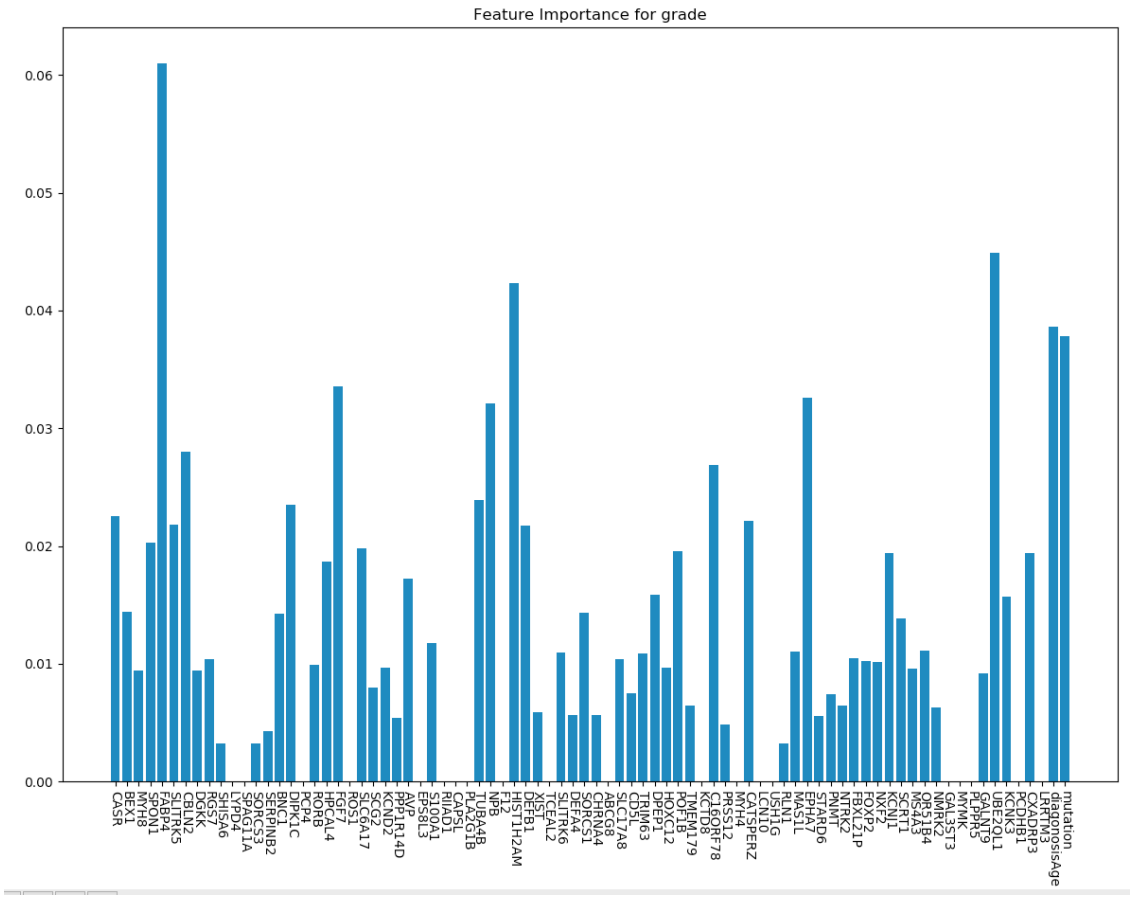
- Another thing is after PCA is applied, original feature names and meanings will get disappeared since high dimension space rotation from PCA. So for this project, it will be impossible to explain what important features are for clinicians. Feature selection will be discussed after.

Feature Selection from Model

(feature selection from scikit learning)

Apply and analysis:

- from sklearn.feature_selection import SelectFromModel
- from sklearn.tree import DecisionTreeClassifier: for category target variables (cancerStage, grade, vitalStatus)
- from sklearn.tree import DecisionTreeRegressor: for numeric target variable (survivalDays)
- Function: *featureSelectionFromModel(patient, mrna, trainingData, type),* Set feature importance threshold ='median'
- The figure shows feature importance for grade target as example.



Total 41 numeric features have been selected (see result)for grade target.

- Including diagnosisAge and mutation, as EDA discussed before.
- Considering pairs of (LYPD4, AG11A), (DGKK, RGS7), (CASR, DGKK), (SPON1, CBLN2) from previous correlation analysis for mrna_data.tsv, only red feature from one pair is selected, another one is not selected because of correlation. Both in pair of (LYPD4, AG11A) are not selected, because they might not be important enough for the target.

```
Selected features from model for  grade : Total numbers is  41
Features are:
 ['CASR' 'BEX1' 'FABP4' 'SLITRK5' 'CBLN2' 'DGKK' 'RGS7' 'DIPK1C' 'HPCAL4'
 'FGF7' 'SLC6A17' 'SCG2' 'PPP1R14D' 'AVP' 'TUBA4B' 'NPB' 'F12' 'HIST1H2AM'
 'DEFB1' 'SLITRK6' 'SORCS1' 'CHRNA4' 'SLC17A8' 'TRIM63' 'DPEP1' 'POF1B'
 'C16ORF78' 'CATSPERZ' 'LCN10' 'MAS1L' 'EPHA7' 'KCNJ1' 'SCRT1' 'MS4A3'
 'OR51B4' 'UBE2QL1' 'KCNK3' 'PCDHB1' 'CXADRP3' 'diagonosisAge' 'mutation']
```

Modeling and Training for Prediction

For best practice of this project, considering non-linear nature and small dataset (deep learning might be the best for the large dataset), ensemble approach might be the best choice (use SVM as baseline model during experiment process). Choose random forest tree as modeling, predict cancerStage, grade, vitalStatus and survivalDays separately.

- Add profiled, informed, race and sex as features for all predictions
- Add cancerStage and grade as features to predict survivalDays
- Add cancerStage, grade and survivalDays as features to predict vitalStatus
- Apply RandomForestClassifier to predict cancerStage, grade, vitalStatus
- Apply RandomForestRegressor to predict survivalDays
- Apply cross validation to evaluate model performance
    - scoring='accuracy' for classifier
    - scoring='neg_mean_squared_error' for regressor
- Try hyper-parameter tuning for n_estimators of random forest

Prediction Results

Includes:

- Selected features from model
- Model performance with n_estimators value
- Confusion matrix for classifier

Grade Prediction

- survivalDays and diagonosisAge selected as features

```
Selected features from model for  grade : Total numbers is  41
Features are:
 ['CASR' 'BEX1' 'FABP4' 'SLITRK5' 'CBLN2' 'DGKK' 'RGS7' 'DIPK1C' 'HPCAL4'
 'FGF7' 'SLC6A17' 'SCG2' 'PPP1R14D' 'AVP' 'TUBA4B' 'NPB' 'F12' 'HIST1H2AM'
 'DEFB1' 'SLITRK6' 'SORCS1' 'CHRNA4' 'SLC17A8' 'TRIM63' 'DPEP1' 'POF1B'
 'C16ORF78' 'CATSPERZ' 'LCN10' 'MAS1L' 'EPHA7' 'KCNJ1' 'SCRT1' 'MS4A3'
 'OR51B4' 'UBE2QL1' 'KCNK3' 'PCDHB1' 'CXADRP3' 'diagonosisAge' 'mutation']
Feature selection: yes
Model parameter of n_estimator is :  10
Predict accuracy of grade 0.4529022764342955
Model parameter of n_estimator is :  20
Predict accuracy of grade 0.4738759662435288
Model parameter of n_estimator is :  30
Predict accuracy of grade 0.4823505425147153
Model parameter of n_estimator is :  40
Predict accuracy of grade 0.48870647471810513
Model parameter of n_estimator is :  50
Predict accuracy of grade 0.5118254024537267
Confusion matrix for Grade:
[[  0  13   0   0   0]
 [  0 169  28   0   0]
 [  0  44 142   0   0]
 [  0  16  58   0   0]
 [  0   5   0   0   0]]
```

CancerStage Prediction

- mutation and diagonosisAge selected as features

```
Selected features from model for  cancerStage : Total numbers is  41
Features are:
 ['CASR' 'MYH8' 'SPON1' 'FABP4' 'CBLN2' 'DGKK' 'RGS7' 'SHISA6' 'SORCS3'
 'SERPINB2' 'HPCAL4' 'SLC6A17' 'SCG2' 'KCND2' 'PPP1R14D' 'PLA2G1B' 'NPB'
 'HIST1H2AM' 'DEFB1' 'XIST' 'TCEAL2' 'SLITRK6' 'SORCS1' 'ABCG8' 'TRIM63'
 'DPEP1' 'POF1B' 'TMEM179' 'KCTD8' 'PRSS12' 'USH1G' 'EPHA7' 'NTRK2'
 'FBXL21P' 'SCRT1' 'NMRK2' 'GALNT9' 'KCNK3' 'PCDHB1' 'diagonosisAge'
 'mutation']
Feature selection: yes
Model parameter of n_estimator is :  10
Predict accuracy of cancer stage 0.3071869347802958
Model parameter of n_estimator is :  20
Predict accuracy of cancer stage 0.3244671418945278
Model parameter of n_estimator is :  30
Predict accuracy of cancer stage 0.32239245309784725
Model parameter of n_estimator is :  40
Predict accuracy of cancer stage 0.322268326417704
Model parameter of n_estimator is :  50
Predict accuracy of cancer stage 0.33050501826435436
Confusion matrix for CancerStage:
[[  0  13   1   0   0   0   0   5   0   0   0]
 [  0 112   1   0   0   0   0   6   0   0   0]
 [  0  45  37   0   0   0   0  17   0   0   0]
 [  0  26   6   2   0   0   0  19   0   0   0]
 [  0   5   1   0   0   0   0   0   0   0   0]
 [  0   1   0   0   0   0   0   3   0   0   0]
 [  0   4   0   0   0   0   0   1   0   0   0]
 [  0  29   2   0   0   0   0  77   0   0   0]
 [  0  16   2   0   0   0   0  31   0   0   0]
 [  0   0   0   0   0   0   0   2   0   0   0]
 [  0   3   2   0   0   0   0   6   0   0   0]]
```

SurvivalDays Prediction

- mutation and diagonosisAge not selected from model, as analysis from EDA

```
Selected features from model for  survivalDays : Total numbers is  41
Features are:
 ['CASR' 'MYH8' 'SPON1' 'SLITRK5' 'CBLN2' 'SORCS3' 'SERPINB2' 'BNC1' 'FGF7'
 'SLC6A17' 'EPS8L3' 'S100A1' 'CAPSL' 'PLA2G1B' 'NPB' 'DEFB1' 'SLITRK6'
 'DEFA4' 'SORCS1' 'CHRNA4' 'SLC17A8' 'CD5L' 'TRIM63' 'POF1B' 'TMEM179'
 'MYH4' 'USH1G' 'RLN1' 'EPHA7' 'STARD6' 'NTRK2' 'FBXL21P' 'FOXP2' 'NXF2'
 'SCRT1' 'NMRK2' 'GAL3ST3' 'PLPPR5' 'GALNT9' 'KCNK3' 'PCDHB1']
Feature selection: yes
Model parameter of n_estimator is :  10
Predict MSE of Survival Days -1.0677661250555475
Model parameter of n_estimator is :  20
Predict MSE of Survival Days -1.052203275372224
Model parameter of n_estimator is :  30
Predict MSE of Survival Days -1.0459327612524547
Model parameter of n_estimator is :  40
Predict MSE of Survival Days -1.0379254682703325
Model parameter of n_estimator is :  50
Predict MSE of Survival Days -1.0395795103146777
```

VitalStatus Prediction

- survivalDays and diagonosisAge selected as features

```
Selected features from model for  vitalStatus : Total numbers is  42
Features are:
 ['BEX1' 'FABP4' 'SLITRK5' 'CBLN2' 'LYPD4' 'SORCS3' 'SERPINB2' 'DIPK1C'
 'HPCAL4' 'FGF7' 'ROS1' 'SLC6A17' 'SCG2' 'KCND2' 'EPS8L3' 'S100A1'
 'RIIAD1' 'CAPSL' 'PLA2G1B' 'NPB' 'XIST' 'TCEAL2' 'DEFA4' 'CHRNA4' 'ABCG8'
 'SLC17A8' 'CD5L' 'DPEP1' 'POF1B' 'PRSS12' 'MYH4' 'USH1G' 'STARD6' 'NTRK2'
 'SCRT1' 'MS4A3' 'NMRK2' 'GALNT9' 'UBE2QL1' 'KCNK3' 'diagonosisAge'
 'survivalDays']
Feature selection: yes
Model parameter of n_estimator is :  10
Predict accuracy of vital status 0.6883842144452719
Model parameter of n_estimator is :  20
Predict accuracy of vital status 0.6905382406127007
Model parameter of n_estimator is :  30
Predict accuracy of vital status 0.7095167180796369
Model parameter of n_estimator is :  40
Predict accuracy of vital status 0.7116086941105556
Model parameter of n_estimator is :  50
Predict accuracy of vital status 0.7179112151189591
Confusion matrix for VitalStatus:
[[306   2]
 [ 99  68]]
```

How to use riskPrediction.py

For example:

```
python riskPrediction.py --rootPath '/Users/tang_li/Desktop/CG/' --predictType 'vitalStatus' --ifSelectFeature ‘yes'
```

Three parameters to run riskPrediction.py

- rootPath: file folder path of all needed data files
- predictType: four options (cancerStage, grade, vitalStatus, survivalDays)
- ifSelectFeature: yes or no, means if do feature selection from model

Note: riskPrediction.py is more research focused, not production ready code, which needs more pipeline work and code optimization. PCA for dimension reduction and figure plot are commented out in riskPrediction.py.

For Questions

1. *What features of the data are most important for QC/QA?*

Please see above

2. *Generally speaking, what are potential sources of ambiguity arising from your approach?*

The dataset of this project is pretty small and un-balanced. Without good dataset, any modeling approaches will be hard to get good performance. Results show categories with large ratio have much better prediction performance. Another thing is about feature selection, there are different ways to try and explain based on different statistics / engineering practice.

3. *What other data might we collect to enhance risk quantification? What quantitative proof do you have?*

The results for categories with small ratio are not good. I think there are three things related with data collection to improve prediction:

- One is collect more data for categories with small ratio.

- Need lifestyle data related with cancer, such as: diet, tobacco, infections, stress, physical activity, environmental pollutants etc
- Need more data for seq, and also association with mrna data

4. *Describe your approach to filing IP claims around your unique classification of risk?*

Please see above

5. *How would you communicate your findings to a clinician?*

I think from my work, data visualization, features directly selected from model, confusion matrix are all good ways to communicate with clinicians.