Hospital Data Set Analysis

Modeling team

About the data set

- Total of 248 patients admitted to the hospital
- For the individuals , total cost to the hospital to be derived

Exploratory Data Analysis - Age

- People aged from 0 to 80 years admitted
- Teenagers [133] are more in number than rest of the age group [115]
- Cost for teenagers is generally low compared to the rest of the admitted population
- Weight and Height of admitted individuals increase with age
- Patients aged more than 20 have hypertension in their medical history
- Teenagers do not have these high end diseases like hypertension , Diabetes in their medical history

Exploratory Data Analysis - Gender

- Total of 166 males admitted as against 82 females
- Average cost to treat a male is generally higher than a female
- Admitted teenaged patients are of similar numbers whereas on the other aged people, most of the admitted patients are male

Exploratory Data Analysis – Martial Status

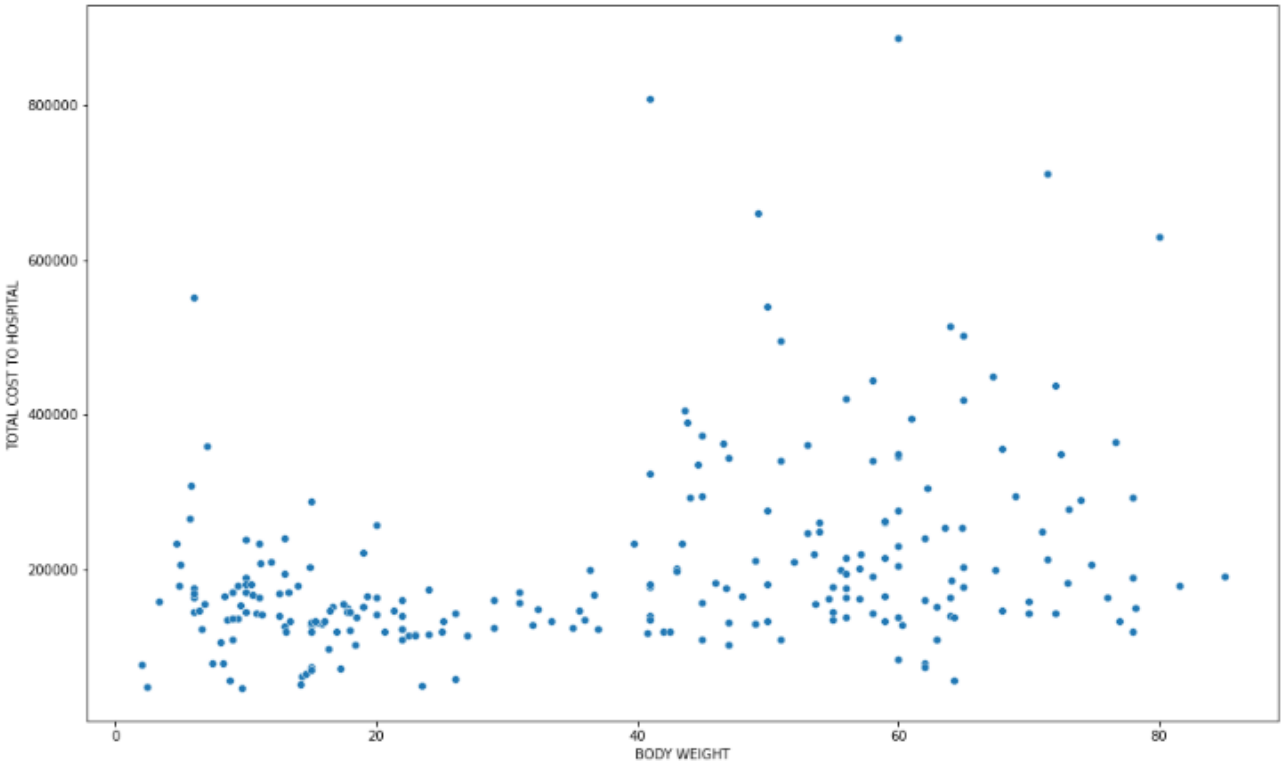
- Out of the 248 admitted patients most of them [140] are unmarried
- On an average the cost to treat a married patient [250975] is generally higher than to treat a unmarried patient [158414]

Exploratory Data Analysis – Key complaints

- Most of the patients admitted were with heart related diseases [55]
- Very few [3] patients had nervous disorders
- The cost for treating a CAD DVD patients is higher than for any other disease in this hospital
- The cost is low for those with general disorders
- Patients of lesser age tend to have ACHD as one of their disease in their medical history

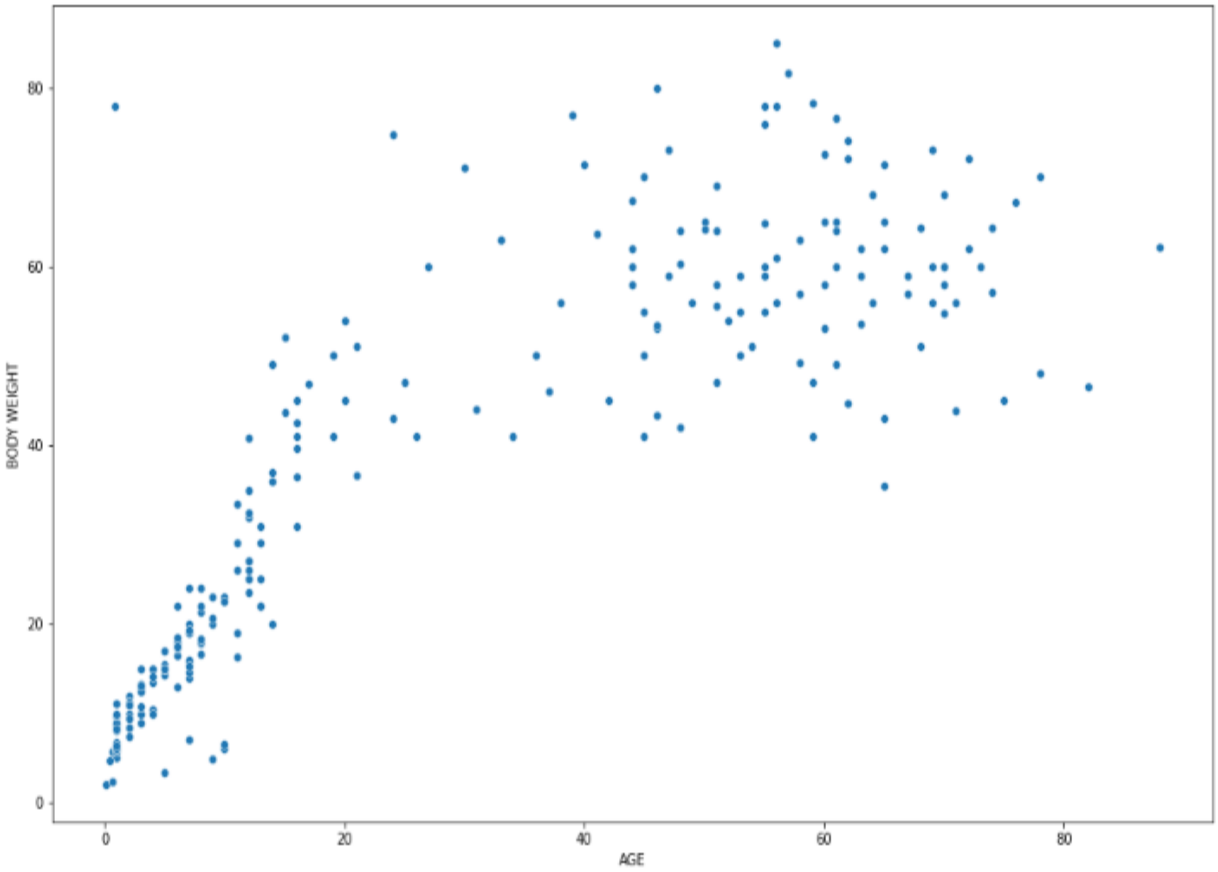
Exploratory Data Analysis – Body Weight

- Weight and cost of treating a patient increase together



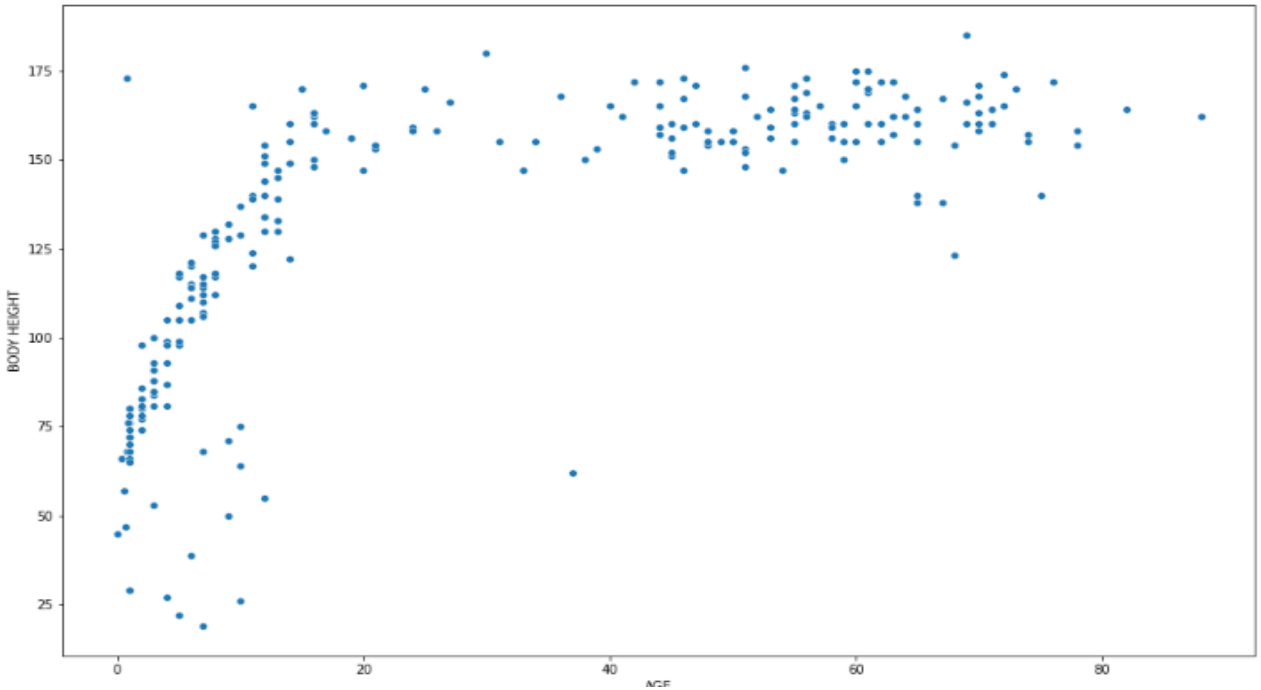
Exploratory Data Analysis – Body Weight - contd

- Weight tend to increase with patients age



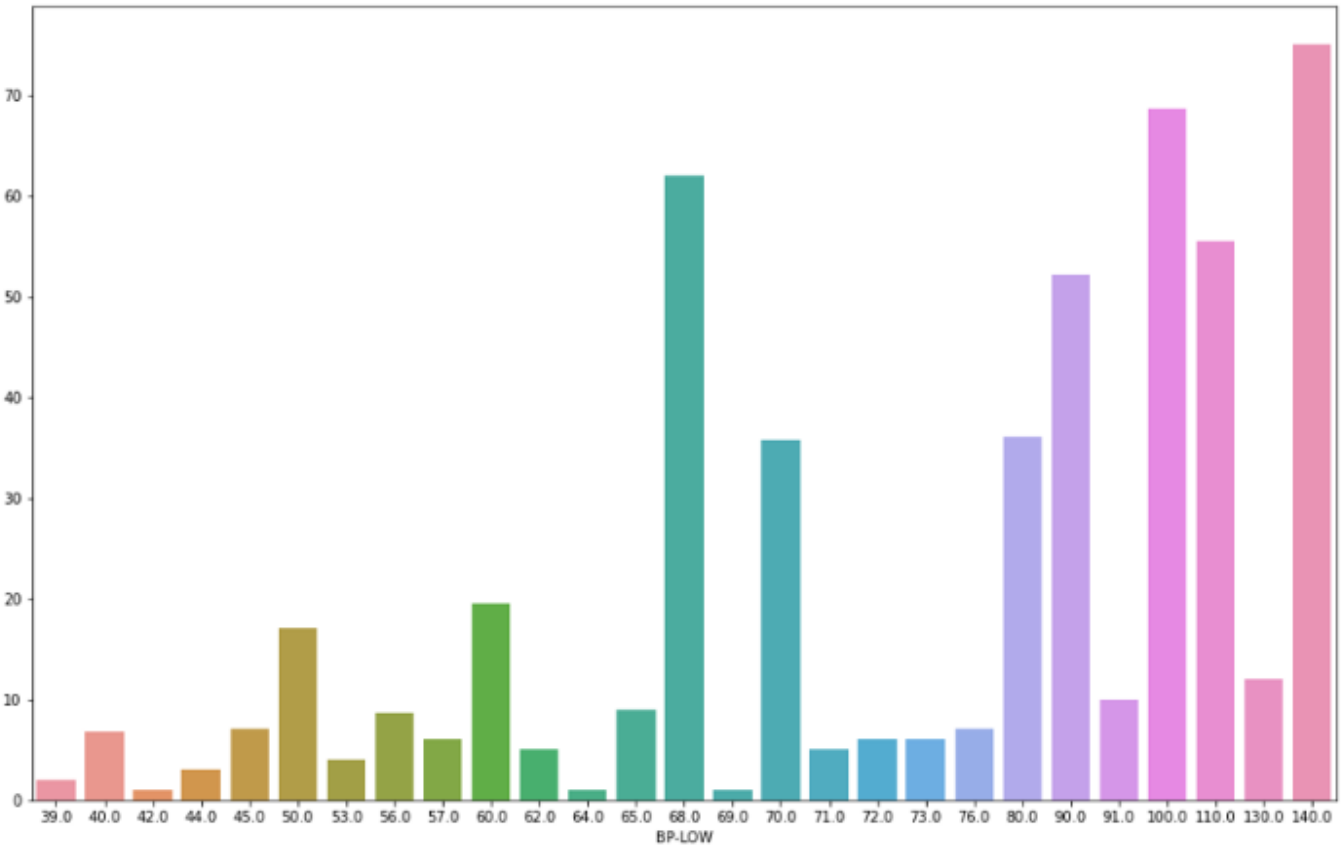
Exploratory Data Analysis - Height

- Height increase with patients age
- The height gets saturated beyond 18 years of age
- Those who are teenagers are around 130 cm range and who are of other aged category tend to be around 150 to 170 cm range



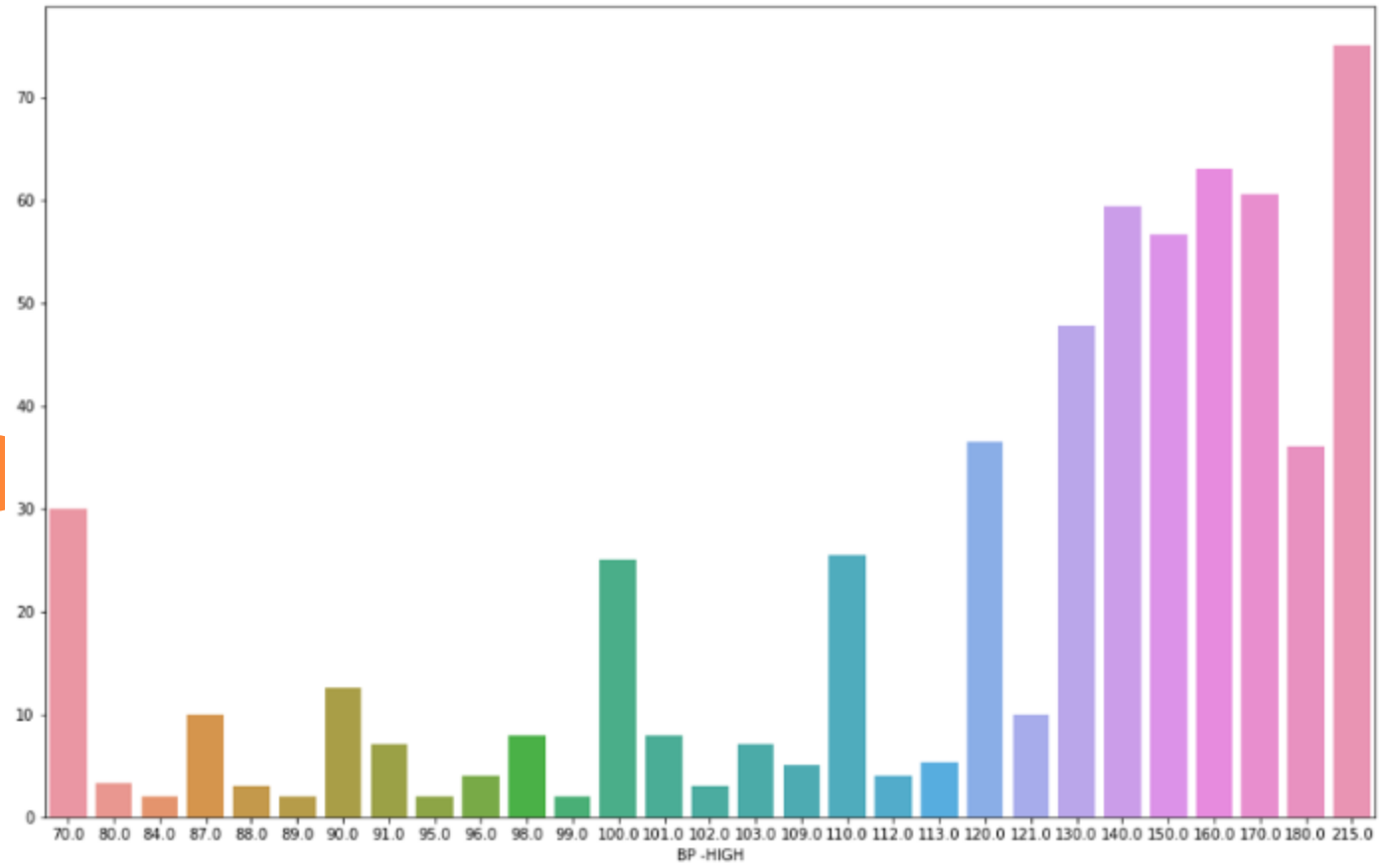
Exploratory Data Analysis – BP -Low

- Those who are of higher aged are with low BP



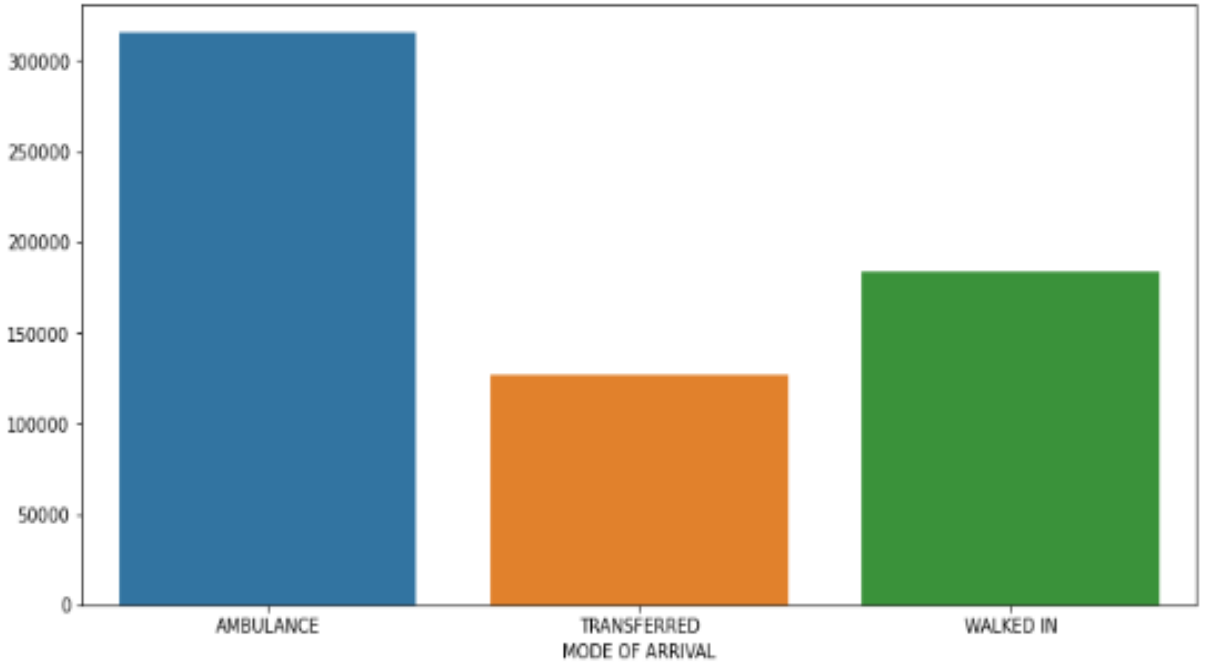
Exploratory Data Analysis - Height

- Those who are of higher aged [50 aged]are with High BP



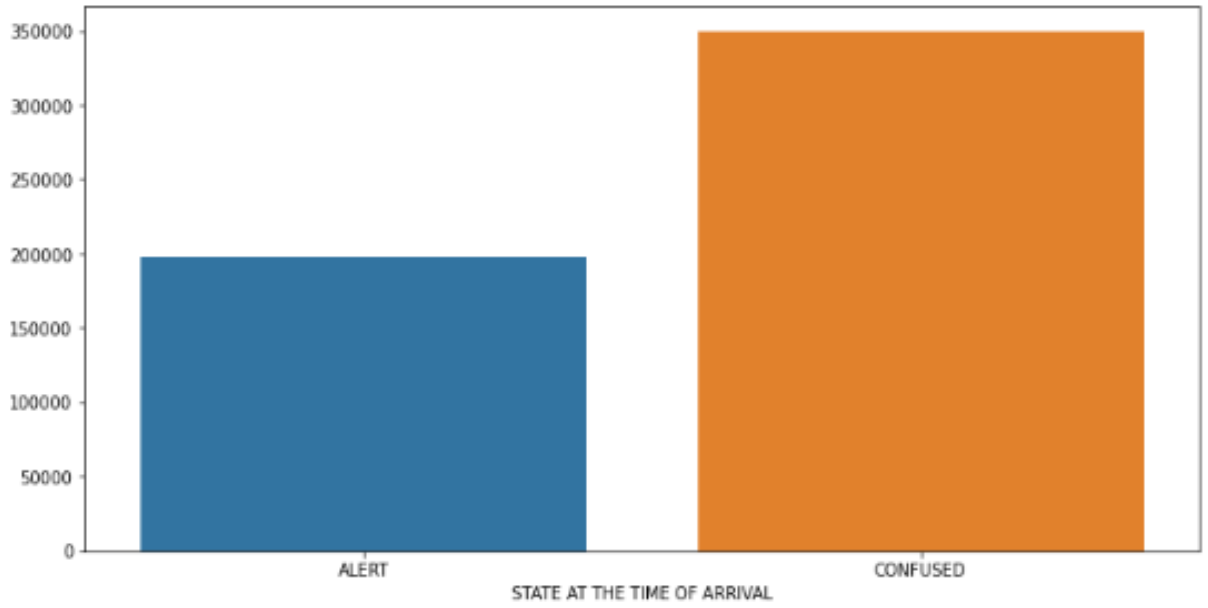
Exploratory Data Analysis – Mode of Arrival

- Those who arrive thru ambulance tend to pay most money



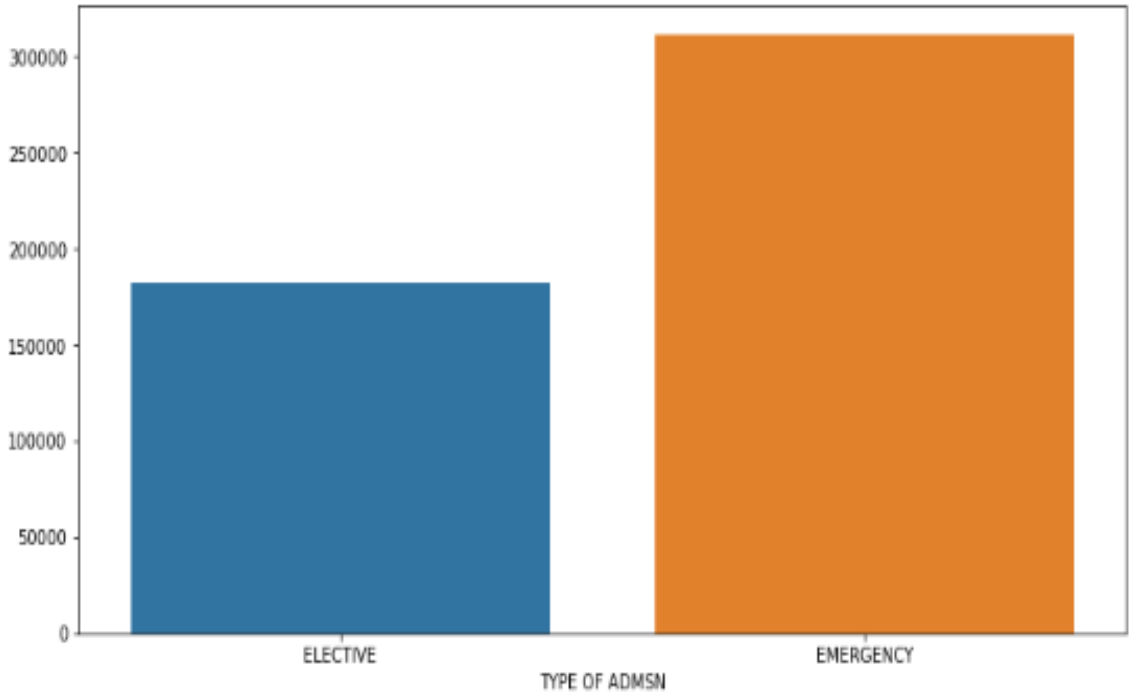
Exploratory Data Analysis – State at the time of arrival

- Most of the patients arrive with a confused state of mind when they get admitted to the hospital [on the nature of disease]



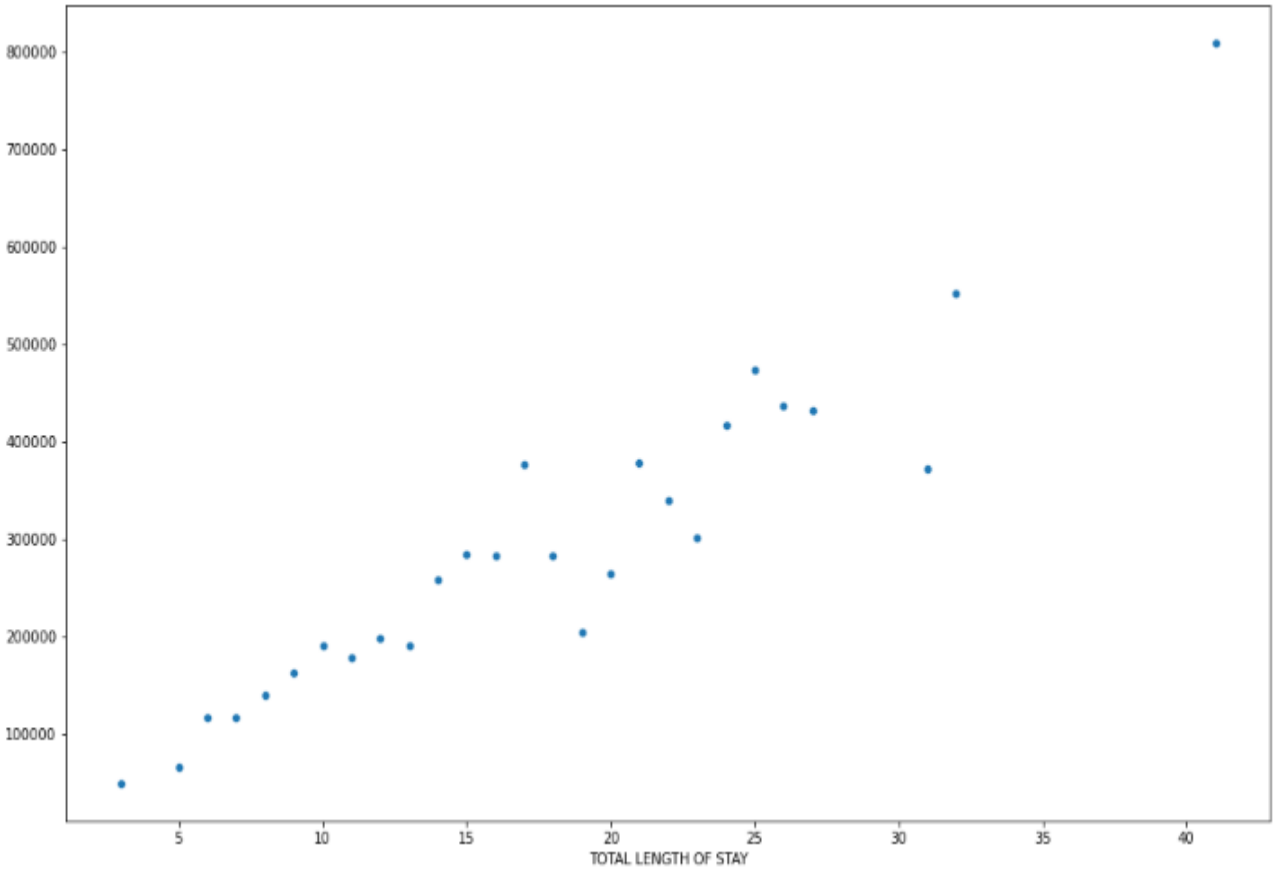
Exploratory Data Analysis – Type of Admission

- Most patients pay more cost when they get admitted in a emergency manner



Exploratory Data Analysis – Length of stay

- The cost of treating a person increase when they had to stay for more days in the hospital

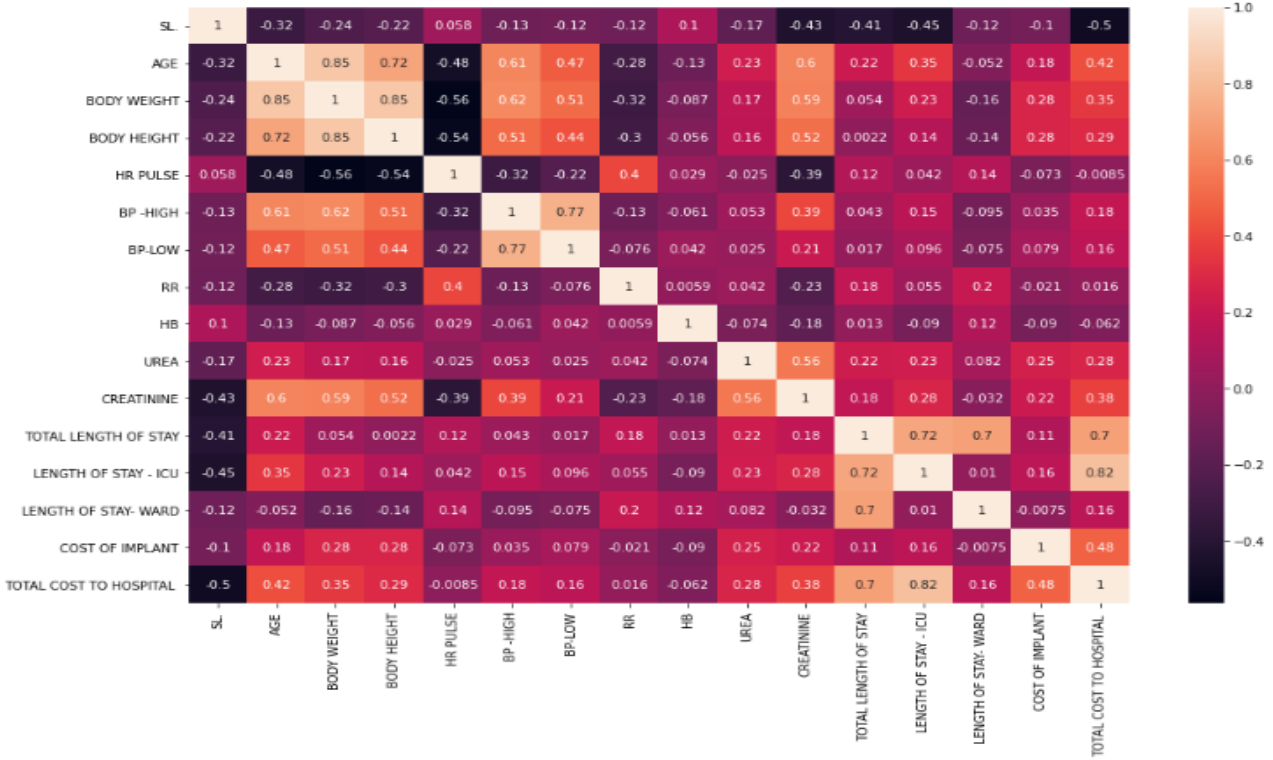


Model Selection

- Out of the 24 columns including SL , there were 7 columns had null values
- Categorical columns with null values were populated with the mode as part of preprocessing
- Numerical columns with null values were populated with the mean as part of preprocessing
- Removed the high correlation columns ['BODY WEIGHT', 'BP-LOW','BODY HEIGHT', 'TOTAL LENGTH OF STAY] as part of preprocessing and running the models with the correlation columns and without correlation columns – Comparisons being done

Model Selection

- Correlation between age and weight/height of the patient body
- Also BP high and low had high correlation
- Total length of stay and length of stay in the ward/ICU have high correlation



Model Selection - Linear Regression

Every column included

Train

MAE: 26256.167142427832
MAPE: 0.128081333455522
MSE: 1846455481.1717098
RSME: 42970.40238549913
R2 Score: 0.8897347739996188

Test

MAE: 35096.23735663518
MAPE: 0.21041611057161083
MSE: 2352443855.9329777
RSME: 48501.998473598775
R2 Score: 0.7800641053903145

Correlation columns removed

Train

MAE: 27563.51799726018
MAPE: 0.13682956923079678
MSE: 2025053055.8889897
RSME: 45000.58950601636
R2 Score: 0.8790694196815042

Test

MAE: 35053.7978566474
MAPE: 0.21575650770126767
MSE: 2155362635.2612066
RSME: 46425.883246969104
R2 Score: 0.7984897245479815

Model Selection - Lasso Regularization

Every column included

Train	Test
MAE: 26262.60358563858	MAE: 35044.43320452568
MAPE: 0.12821279573659863	MAPE: 0.2096941093280193 5
MSE: 1846581479.661422	MSE: 2346477866.507358
RSME: 42971.8684683529	RSME: 48440.45691885408
R2 Score: 0.8897272497175087	R2 Score: .7806218807515612

Correlation columns removed

Train	Test
MAE: 27555.358844066857	MAE: 35071.21639499769
MAPE: 0.13679691082474724	MAPE: 0.21591516899076357
MSE: 2025097335.4925623	MSE: 2157597549.654171
RSME: 45001.08149247707	RSME: 46449.94671314673
R2 Score: 0.8790667754257695	R2 Score: 0.7982807767785574

Model Selection - Ridge Regularization

Every column included

Train

MAE: 26804.08813463882
MAPE: 0.13487977654797043
MSE: 1906838671.7389975
RSME: 43667.36392019786 R2
Score: 0.886128856487705

Test

MAE: 29795.32131455586
MAPE: 0.17452823512722235
MSE: 1723035426.266261
RSME: 41509.461888420825 R2
Score: 0.8389090830098657

Correlation columns removed

Train

MAE: 27525.729501719048
MAPE: 0.1368701324349566 6
MSE: 2026023153.24726
RSME: 45011.36693377863
R2 Score: 0.8790114881442737

Test

MAE: 34699.248496555025
MAPE: 0.2132951058615914
MSE: 2113776837.133686
RSME: 45975.828835744614
R2 Score: 0.8023776854407216

Model Selection - Elastic Net

Every column included

Train	Test
MAE: 29318.867015399857	MAE: 25415.479933924056
MAPE: 0.15750647631241319	MAPE: 0.1653575461106671
MSE: 2491318516.225202	MSE: 1072703009.3931755
RSME: 49913.10966294529	RSME: 32752.14511132325
R2 Score: 0.8512253330601907	R2 Score: 0.8997102968360442

Correlation columns removed

Train	Test
MAE: 32538.783272762335	MAE: 35226.113554730735
MAPE: 0.1746730610733512	MAPE: 0.23738703148897686
MSE: 3133057676.572782	MSE: 2204019106.1940684
RSME: 55973.72309014992	RSME: 46946.981864589216
R2 Score: 0.812902441297797	R2 Score: 0.7939407086655493

Model Selection - SVM

Every column included

Train

MAE: 76874.23567036873
MAPE: 0.3648245199089476
MSE: 18241391277.902863
RSME: 135060.69479275923
R2 Score: -0.089325549589188 92

Test

MAE: 70676.25175895492
MAPE: 0.39147315185924886
MSE: 11230081361.1446 RSME: 105972.07821470994
R2 Score: -0.04992856023904224

Correlation columns removed

Train

MAE: 76878.24145249224
MAPE: 0.3648293063644489
MSE: 18242784541.62672
RSME: 135065.85261133444
R2 Score: -0.08940875145403537

Test

MAE: 70678.67878889255
MAPE: 0.3914700009902881
MSE: 11231512263.34738
RSME: 105978.82931674315
R2 Score: -.05006233888599021

Model Selection - KNN

Every column included

Train	Test
MAE: 42797.767537572254 MAPE: 0.2149823423206306 5 MSE: 5328105363.734529 RSME: 72993.87209714613 R2 Score: 0.6818202507036795	MAE: 43627.9124 MAPE: 0.2764556605947345 3 MSE: 4244050252.1363964 RSME: 65146.375587106886 R2 Score: 0.6032130643126927

Correlation columns removed

Train	Test
MAE: 47655.36686705203 MAPE: 0.2376451722709426 MSE: 6451573037.317036 RSME: 80321.68472658574 R2 Score: 0.6147298614715 02	MAE: 52982.92477333333 MAPE: 0.3030638086573779 MSE: 6225800613.204821 RSME: 78903.74270720509 R2 Score: 0.4179342371664 4806

Model Selection - **Decision tree**

Every column included

Train	Test
MAE: 0.0	MAE: 44669.26093333333
MAPE: 0.0	MAPE: 0.277875199637085 57
MSE: 0.0	MSE: 4490848108.513907
RSME: 0.0	RSME: 67013.79043535671
R2 Score: 1.0	R2 Score: 0.5801393118007062

- Due to over fitting issue, the decision tree model is flawed and not recommended for this one

Model Selection - **Bagging Regressor**

Every column included

Train	Test
MAE: 14596.536647398843	MAE: 30398.641413333335
MAPE: 0.078648269667703 14	MAPE: 0.20188154883520182
MSE: 746537439.2375195	MSE: 1852635420.2599647
RSME: 27322.83732040872 R2	RSME: 43042.251570520384 R2
Score: 0.955418844215494	Score: 0.8267924535104931

Model Selection - **Random Forest Regressor**

Every column included

Train	Test
MAE: 13440.977246820808	MAE: 32909.287109333345
MAPE: 0.06922732353108271	MAPE: 0.21384077296416487
MSE: 589964640.6800026	MSE: 2234638190.0356545
RSME: 24289.187731993068 R2	RSME: 47271.95987089656 R2
Score: 0.9647689396792098	Score: 0.7910780534825812

Model Selection - **Ada Boost Regressor**

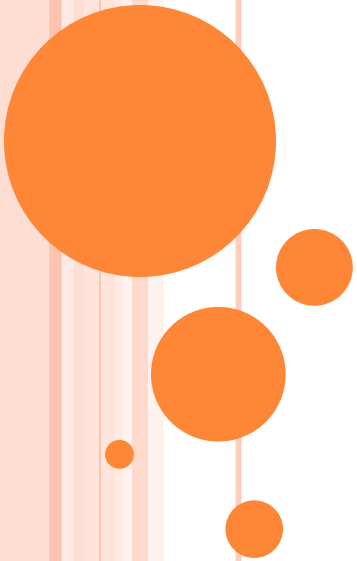
Every column included

Train	Test
MAE: 21734.091386477772	MAE: 32488.379714565766
MAPE: 0.1487455879841437	MAPE: 0.20470661480585983
MSE: 725704512.4234257	MSE: 3422175427.895732
RSME: 26938.903326294218 R2	RSME: 58499.36262811529 R2
Score: 0.956662929159042	Score: 0.6800522093875745

Model Selection - **Gradient Boosting Regressor**

Every column included

Train	Test
MAE: 7499.701111934845	MAE: 29782.342357776277
MAPE: 0.05080993315118985	MAPE: 0.1911274596999689
MSE: 88871251.01753236	MSE: 2448275328.707087
RSME: 9427.154980031482 R2	RSME: 49480.04980501826 R2
Score: 0.9946928541314376	Score: 0.7711045798980564



Model Selection - **LGBM Regressor**

Every column included

Train	Test
MAE: 21147.96099587732	MAE: 38984.9577923248
MAPE: 0.1056291513746657 9	MAPE: 0.2350998676228587 3
MSE: 1500323984.3477814	MSE: 3436731299.717175
RSME: 38734.01585619262 R2	RSME: 58623.64113322521 R2
Score: 0.9104047918323376	Score: 0.6786913443098389

Model Selection - **XGB Regressor**

Every column included

Train	Test
MAE: 2.1752420520236915	MAE: 35851.38492708333
MAPE: 1.5054729208686 49e-05	MAPE: 0.22870442802489174
MSE: 11.725865868861018	MSE: 2513438951.3613696
RSME: 3.4243051658491273 R2	RSME: 50134.209392004675 R2
Score: 0.9999999992997637	Score: 0.7650122688708103

Model Selection - **Cat Boost Regressor**

Every column included

Train	Test
MAE: 1963.9701903691846	MAE: 25695.507408215082
MAPE: 0.013276526628004014	MAPE: 0.1912796357219781 2
MSE: 6171792.871333245	MSE: 1311035382.737259
RSME: 2484.3093348722186 R2	RSME: 36208.22258461825 R2
Score: 0.999631437560924	Score: 0.8774280036311797

Model Selection - **GRid Search CV**

Every column included

Train	Test
MAE: 8609.146636851898	MAE: 27715.24518511961
MAPE: 0.0575249029255634	MAPE: 0.19710416670894607
MSE: 113399539.0951359	MSE: 1566984255.2572997
RSME: 10648.921968684712 R2	RSME: 39585.15195445509
Score: 0.993228092453802	R2 Score: 0.853498699596967 2

Model Selection - **Polynomial Regression**

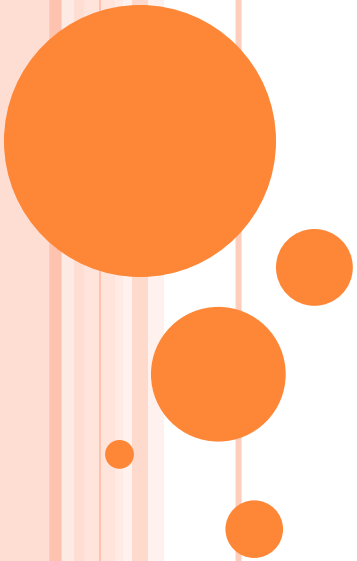
Every column included

Train	Test
MAE: 3.8427991168088994e-10	MAE: 70900.91496628764
MAPE: 2.4650141074949133e-15	MAPE: 0.42712037110745454
MSE: 4.554604484576882e-19	MSE: 10975507056.037888
RSME: 6.748780989613518e-10	RSME: 104764.05421726427 R2
R2 Score: 1.0	Score: -0.026127767970578 9

Model Selection - **Voting Regressor**

Every column included

Train	Test
MAE: 15241.12064150661	MAE: 23013.55227577516
MAPE: 0.08316358021390366	MAPE: 0.16600473114763808
MSE: 648344014.6518197	MSE: 951896830.2954404
RSME: 25462.600312061997 R2	RSME: 30852.82532111833 R2
Score: 0.9612826845648009	Score: 0.9110047704564149



Model Selection - **Stacking CV Regressor**

Every column included

Train	Test
MAE: 18369.91830860198	MAE: 28560.794858746987
MAPE: 0.092241407886283 17	MAPE: 0.18116965770626964
MSE: 762399625.8286237	MSE: 1406262604.131735
RSME: 27611.584993053617 R2	RSME: 37500.168054713235 R2
Score: 0.9544715982043319	Score: 0.8685249711206409

Model Selection - **Least Angle Regression**

Every column included

Train	Test
MAE: 27524.438034682084	MAE: 28171.703866666667
MAPE: 0.1380471857512925	MAPE: 0.1687909273096195
MSE: 2037834618.7018185	MSE: 1667541514.278348
RSME: 45142.38162416576 R2	RSME: 40835.54229195871 R2
Score: 0.8783061400213306	Score: 0.8440973484588671

- Least Angle Regression or LARS for short provides an alternate, efficient way of fitting a Lasso regularized regression model that does not require any hyper parameters.
- LARS averages the attributes and proceeds in a direction that is at the same angle to the attributes.
- Find a variable that is most highly correlated to the residual. Move the regression line in this direction until we reach another variable that has the same or higher correlation.

Model Selection - **Orthogonal Matching Pursuit**

Every column included

Train	Test
MAE: 31195.989377258775	MAE: 27999.7018266594
MAPE: 0.1710781954863162	MAPE: 0.188427828657618 98
MSE: 2656793180.105319	MSE: 1686697813.1597161
RSME: 51544.089671904374 R2	RSME: 41069.426744960976 R2
Score: 0.8413436427643056	Score: 0.8423063778810748

- Matching pursuit (MP) is a sparse approximation algorithm which finds the "best matching" projections of multidimensional data onto the span of an over-complete

- The orthogonal matching pursuit (OMP) [79] or orthogonal greedy algorithm is more complicated than MP

Model Selection - **Bayesian Ridge**

Every column included

Train	Test
MAE: 70514.32818216766	MAE: 66552.06967169991
MAPE: 0.3962061277036637 6	MAPE: 0.4269480595636386
MSE: 13006413885.308813	MSE: 7989502566.273995
RSME: 114045.66578923029 R2	RSME: 89384.01739838054 R2
Score: 0.2232928542593412	Score: 0.2530413042725078

- Matching pursuit (MP) is a sparse approximation algorithm which finds the "best matching" projections of multidimensional data onto the span of an over-complete

Model Selection – Conclusion

- RSME: 36208.22258461825 for Cat boost regressor
- RSME: 32752.14511132325 for elastic net
- RSME: 30852.82532111833 for Voting regressor

Performance tuning:

Cat boost and elastic net gave good result and hence we combined it using Voting regressor

Thank you!