|  |  |  |
| --- | --- | --- |
| Image result for insofe | Research Internship | |
|  | |  |

**DIABETES**

(130-US hospitals for years 1999-2008)

**Karthik Reddy Mathuru**

**Batch-31**

**Data:**

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals.  It includes over 50 features and 101766 records representing patient and hospital outcomes.

**Objective:**

To classify whether a patient is getting readmitted or not with respect to various factors.

**Diabetes:**

**Diabetes** is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. We have two types

**Causes:**

* **Type 1 Diabetes:** Glucose comes from the foods you eat**.** Pancreas makes a hormone called insulin. It's what lets your cells turn glucose from the food you eat into energy. Prediabetes (Type 1), their blood glucose is not normal, but not high enough to be diabetes yet**.**
* **Type 2 Diabetes:** People with “type 2 diabetes“ make insulin, but their cells don't use it as well as they should. At first, the pancreas makes more insulin to try to get glucose into the cells. But eventually it can't keep up, and the sugar builds up in your blood instead.

**Attribute Information:**

Feature name Type Description and values % missing

Encounter ID Numeric Unique identifier of an encounter 0%

Patient number Numeric Unique identifier of a patient 0%

Race Nominal Values: Caucasian, Asian, African American, 2%

Hispanic, and other

Gender Nominal Values: male, female, and unknown/invalid 0%

Age Nominal Grouped in 10-year intervals: 0% 0, 10), 10, 20), …, 90, 100)

Weight Numeric Weight in pounds. 97%

Admission type Nominal Integer identifier corresponding to 9 distinct 0%

values, for example, emergency, urgent, elective,

newborn, and not available

Discharge disposition Nominal Integer identifier corresponding to 29 distinct values, 0%

for example, discharged to home, expired, and

not available

Admission source Nominal Integer identifier corresponding to 21 distinct values, 0%

for example, physician referral, emergency room,

and transfer from a hospital

Time in hospital Numeric Integer number of days between admission 0%

and discharge

Payer code Nominal Integer identifier corresponding to 23 distinct values, 52%

for example, Blue Cross/Blue Shield, Medicare,

and self-pay

Medical specialty Nominal Integer identifier of a specialty of the admitting 53%

physician, corresponding to 84 distinct values, for

example, cardiology, internal medicine, family/

general practice, and surgeon

Number of lab Numeric Number of lab tests performed during the 0%

Procedures encounter

Number of Numeric Number of procedures (other than lab tests) 0%

Procedures performed during the encounter

Number of Numeric Number of distinct generic names administered 0%

Medications during the encounter

Number of Numeric Number of outpatient visits of the patient in the 0%

outpatient visits year preceding the encounter

Number of Numeric Number of emergency visits of the patient in the 0%

emergency visits year preceding the encounter

Number of Numeric Number of inpatient visits of the patient in the 0%

inpatient visits year preceding the encounter

Diagnosis 1 Nominal The primary diagnosis (coded as first three 0% digits of ICD9); 848 distinct values. (International Classification of diseases)

Diagnosis 2 Nominal Secondary diagnosis (coded as first three digits 0% of ICD9); 923 distinct values

Diagnosis 3 Nominal Additional secondary diagnosis (coded as first 1% three digits of ICD9); 954 distinct values

Number of Numeric Number of diagnoses entered to the system 0%

diagnoses

Glucose serum Nominal Indicates the range of the result or if the test was 0%

test result not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured

A1c test result Nominal Indicates the range of the result or if the test was 0% not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.

Change of Nominal Indicates if there was a change in diabetic 0%

medications medications (either dosage or generic name). Values: “change” and “no change”

Diabetes Nominal Indicates if there was any diabetic medication 0%

Medications prescribed. Values: “yes” and “no”

23 features for Nominal For the generic names: metformin, repaglinide, 0%

medications nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed

Readmitted Nominal Days to inpatient readmission. Values: “<30” if the 0% patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.

**Glucose serum test result:**

* Fasting blood sugar (FBS):

A **fasting blood sugar** (**FBS**) **level** is the result of a **blood** sample taken after a patient fasts for at least eight hours. ... Pre diabetes is a **fasting blood sugar level** from 100 to 125 mg/dL, or HbA1C of 5.7%-6.4%. Diabetes is a **fasting blood sugar level** of 126 mg/dL and above, or HbA1C of 6.5% or higher.

* Fasting plasma glucose (FPG): 8 or 12 or 14 hours after eating:

Normal fasting blood glucose -- or blood sugar -- is between 70 and 100 milligrams per deciliter or mg/dL for people who do not have diabetes. The standard diagnosis of diabetes is made when two separate blood tests show that your fasting blood glucose level is greater than or equal to 126 mg/dL.

* Glucose tolerance test: continuous testing. If you're being tested for type-2 diabetes, two hours after drinking the glucose solution: A normal blood glucose level is lower than 140 mg/dL (7.8 mmol/L). A blood glucose level between 140 mg/dL and 199 mg/dL (7.8 and 11 mmol/L) is considered impaired glucose tolerance, or pre diabetes.
* Postprandial glucose test (PC): 2 hours after eating

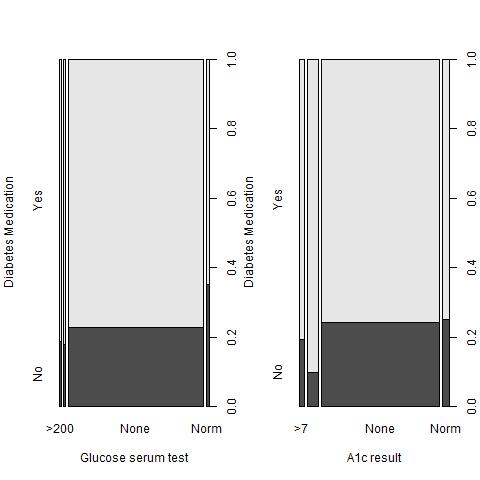
For someone without diabetes, a fasting blood sugar on awakening should be under 100 mg/dl. Before-meal normal sugars are 70–99 mg/dl. “Postprandial” sugars taken two hours after meals should be less than 140 mg/dl. ... There is also a long-term glucose test called a hemoglobin A1c, HbA1c, or just A1C.

* Random glucose test.

The reference values for a "normal" random glucose test in an average adult are 79–160 mg/dl (4.4–7.8 160–200 mg/dl is considered pre-diabetes, and > 200 mg/dl is considered diabetes according to ADA guidelines (you should visit your doctor or a clinic for additional tests however as a random glucose.

**A1C test**-

The **A1C test** is a blood **test** that provides information about a person's average levels of blood glucose, also called blood sugar, over the past 3 months. The **A1C test** is sometimes called the hemoglobin **A1c**, HbA1c, or glycohemoglobin **test**. An **A1C** between 5.7 and 6.4 percent signals pre diabetes. Type 2 diabetes is diagnosed when the**A1C** is over 6.5 percent.



**Diagnosis:**

We have list of codes which are **International Statistical Classification of Diseases and Related Health Problems.** These codes tell’s us which diagnosis is done for a particular patient and which kind of problem he is facing with.

* [List of ICD-9 codes 001–139: infectious and parasitic diseases](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_001%E2%80%93139:_infectious_and_parasitic_diseases)
* [List of ICD-9 codes 140–239: neoplasms](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_140%E2%80%93239:_neoplasms)
* [List of ICD-9 codes 240–279: endocrine, nutritional and metabolic diseases, and immunity disorders](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_240%E2%80%93279:_endocrine,_nutritional_and_metabolic_diseases,_and_immunity_disorders)
* [List of ICD-9 codes 280–289: diseases of the blood and blood-forming organs](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_280%E2%80%93289:_diseases_of_the_blood_and_blood-forming_organs)
* [List of ICD-9 codes 290–319: mental disorders](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_290%E2%80%93319:_mental_disorders)
* [List of ICD-9 codes 320–389: diseases of the nervous system and sense organs](https://en.wikipedia.org/w/index.php?title=List_of_ICD-9_codes_320%E2%80%93389:_diseases_of_the_nervous_system_and_sense_organs&action=edit&redlink=1)
* [List of ICD-9 codes 390–459: diseases of the circulatory system](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_390%E2%80%93459:_diseases_of_the_circulatory_system)
* [List of ICD-9 codes 460–519: diseases of the respiratory system](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_460%E2%80%93519:_diseases_of_the_respiratory_system)
* [List of ICD-9 codes 520–579: diseases of the digestive system](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_520%E2%80%93579:_diseases_of_the_digestive_system)
* [List of ICD-9 codes 580–629: diseases of the genitourinary system](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_580%E2%80%93629:_diseases_of_the_genitourinary_system)
* [List of ICD-9 codes 630–679: complications of pregnancy, childbirth, and the puerperium](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_630%E2%80%93679:_complications_of_pregnancy,_childbirth,_and_the_puerperium)
* [List of ICD-9 codes 680–709: diseases of the skin and subcutaneous tissue](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_680%E2%80%93709:_diseases_of_the_skin_and_subcutaneous_tissue)
* [List of ICD-9 codes 710–739: diseases of the musculoskeletal system and connective tissue](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_710%E2%80%93739:_diseases_of_the_musculoskeletal_system_and_connective_tissue)
* [List of ICD-9 codes 740–759: congenital anomalies](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_740%E2%80%93759:_congenital_anomalies)
* [List of ICD-9 codes 760–779: certain conditions originating in the perinatal period](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_760%E2%80%93779:_certain_conditions_originating_in_the_perinatal_period)
* [List of ICD-9 codes 780–799: symptoms, signs, and ill-defined conditions](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_780%E2%80%93799:_symptoms,_signs,_and_ill-defined_conditions)
* [List of ICD-9 codes 800–999: injury and poisoning](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_800%E2%80%93999:_injury_and_poisoning)
* [List of ICD-9 codes E and V codes: external causes of injury and supplemental classification](https://en.wikipedia.org/wiki/List_of_ICD-9_codes_E_and_V_codes:_external_causes_of_injury_and_supplemental_classification)

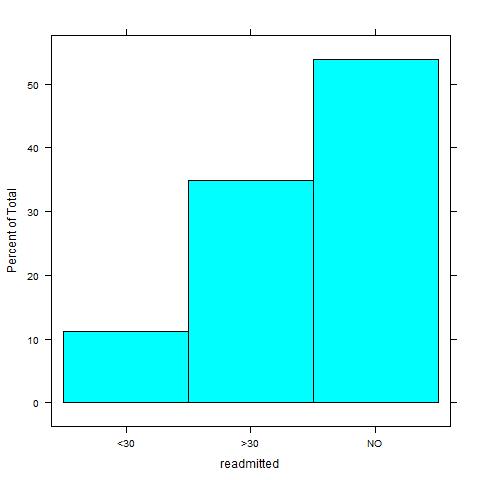
**Class Imbalance:**

There is a class imbalance.

* “What happens if we have class imbalance?”

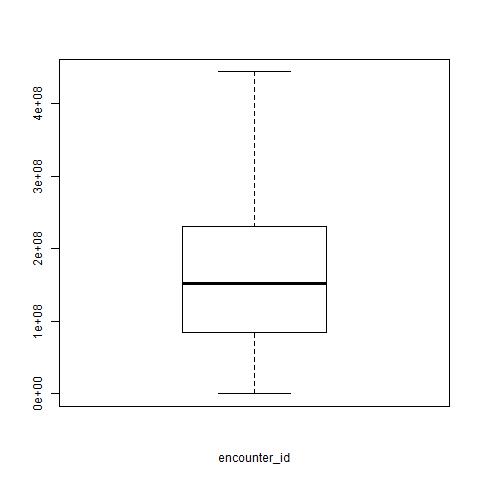
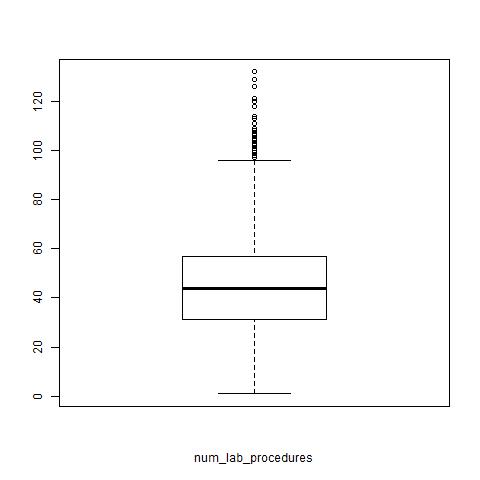
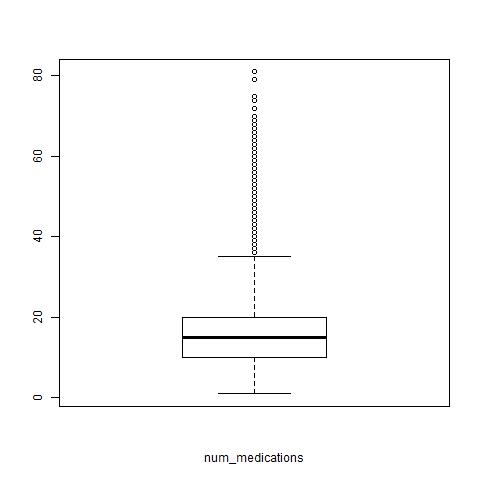
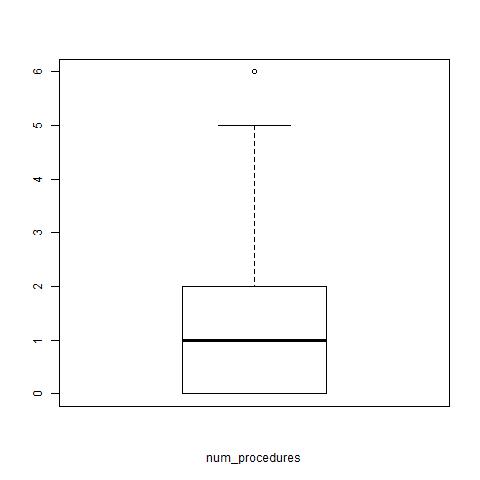
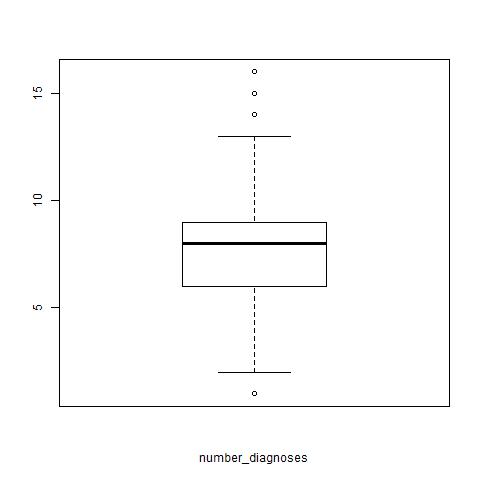
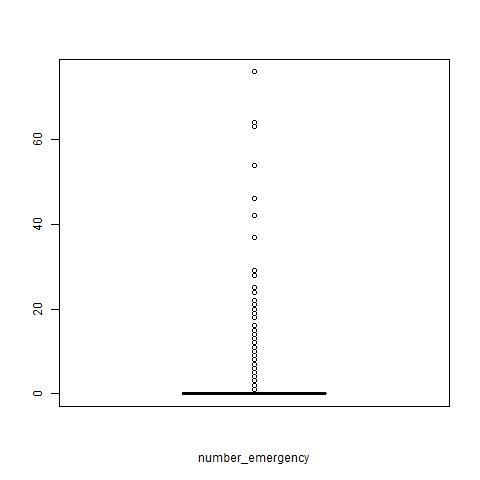
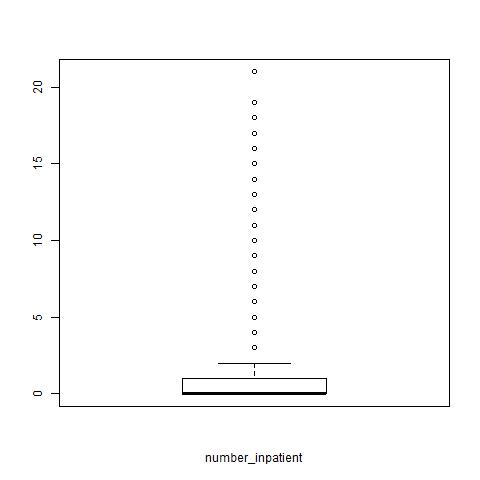
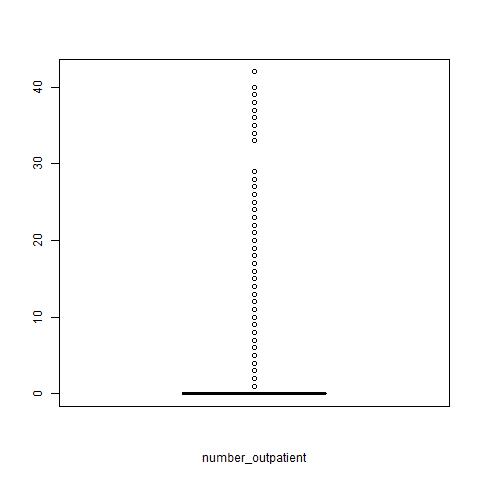
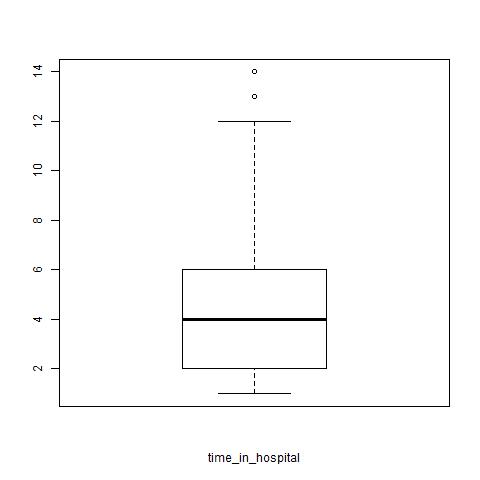
The model which is built will be biased to majority class and will not predict the minority class when it comes to test data.

* we use SMOTE (synthetic minority over sampling Technique) which over samples minority class (which is ‘<30 days’ in our case).

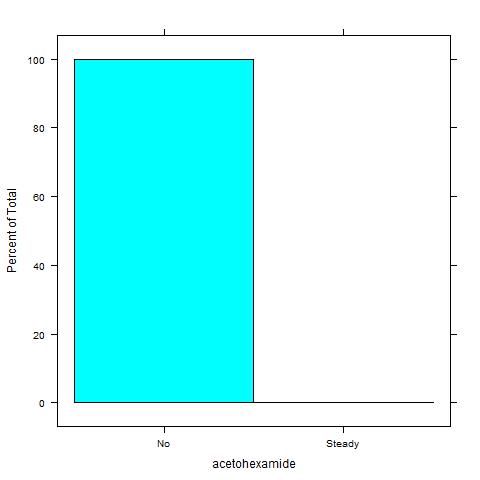
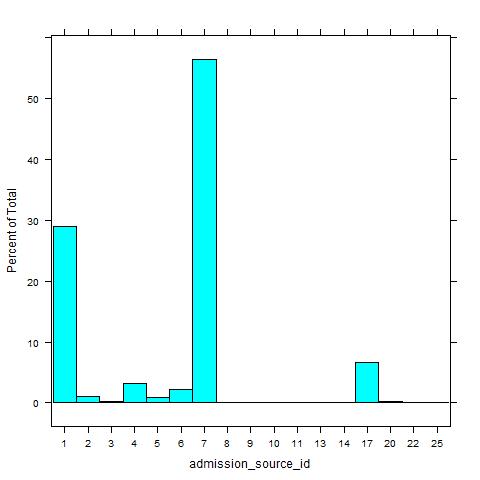
****

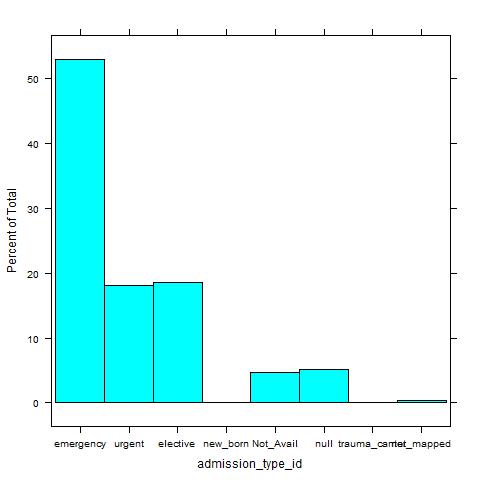
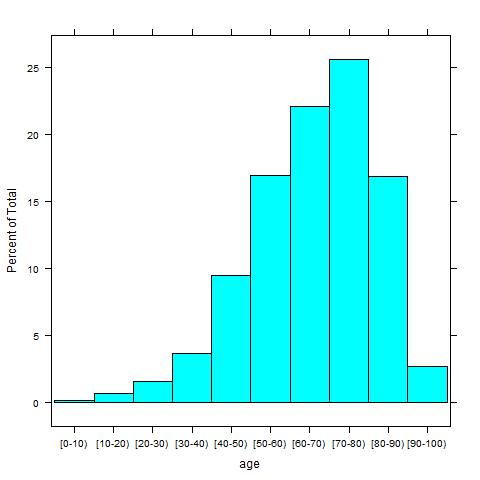
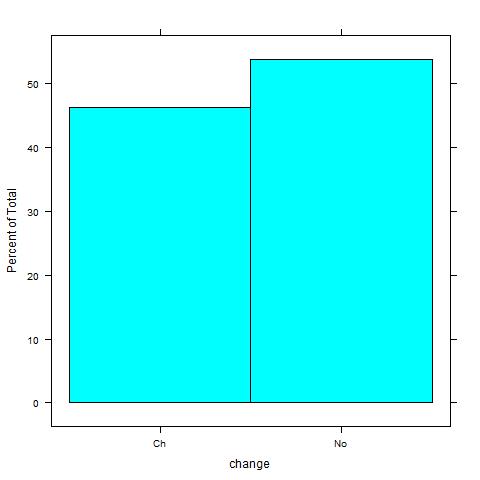
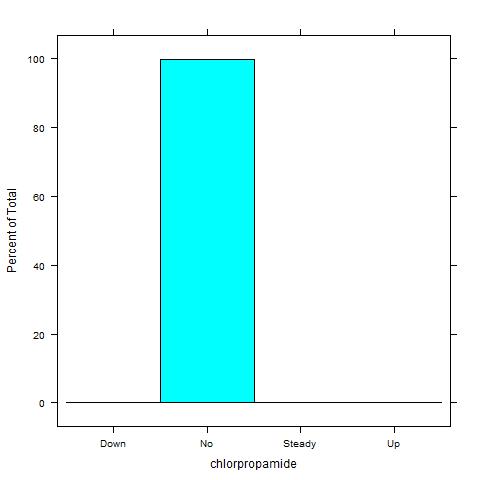
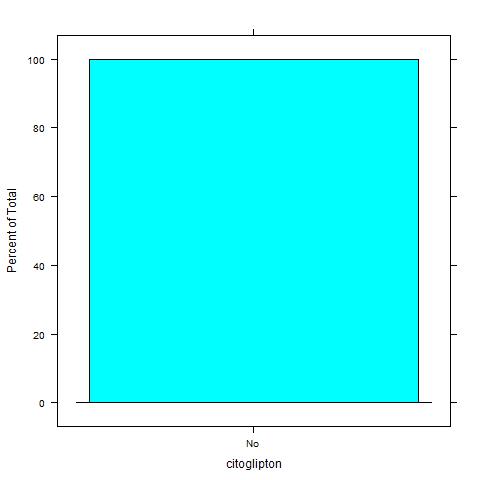
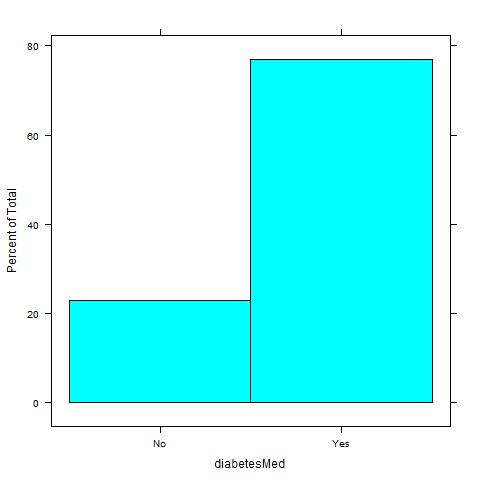
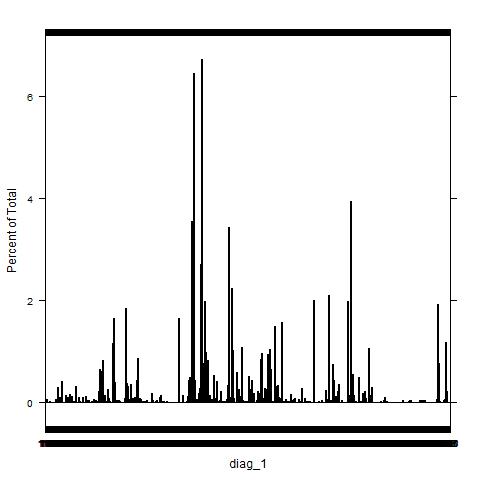
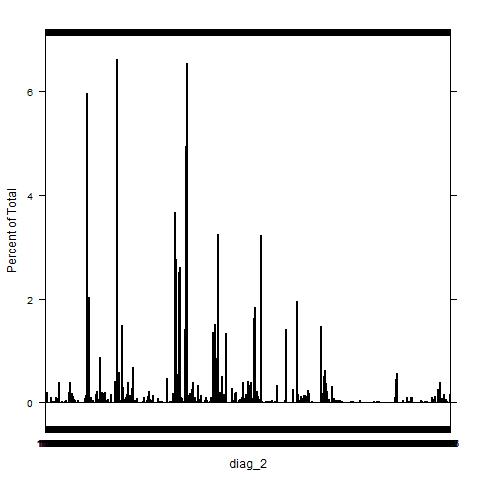
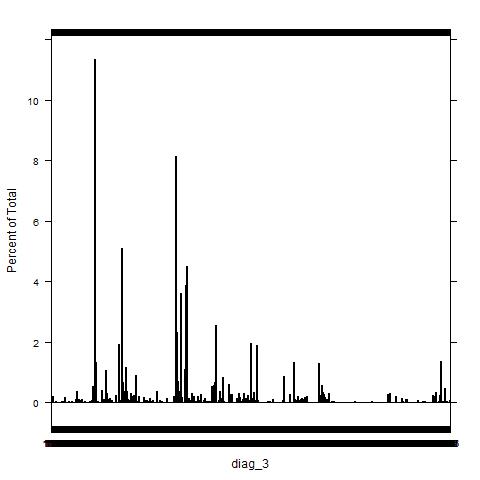
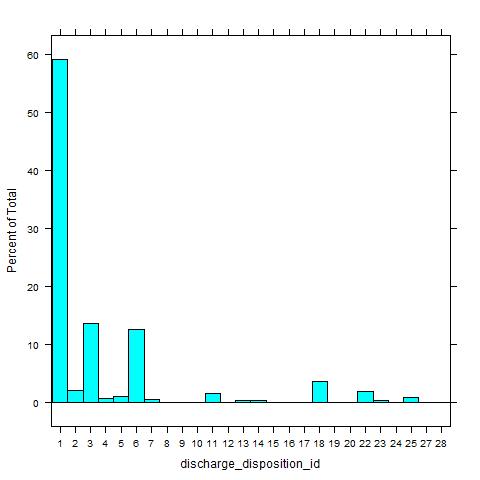
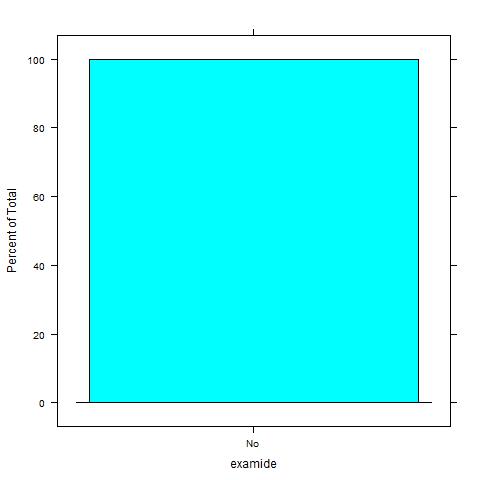
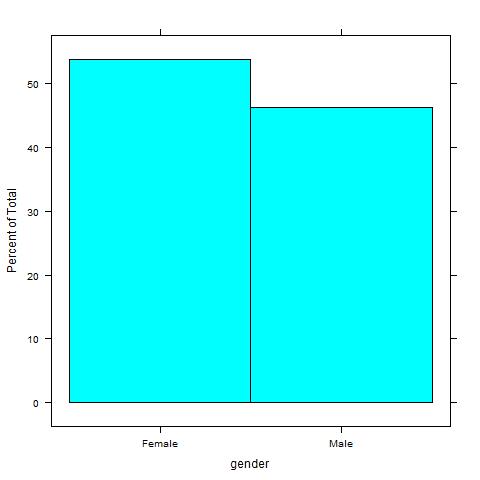
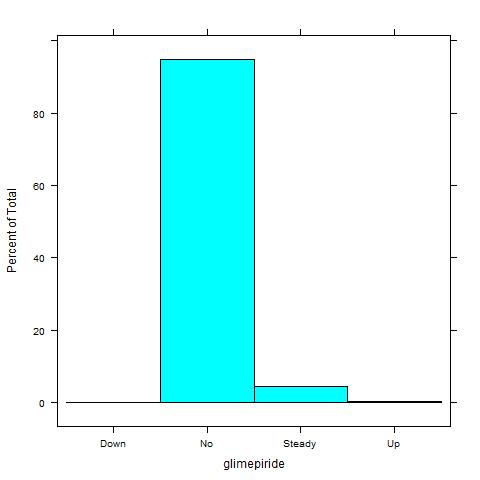
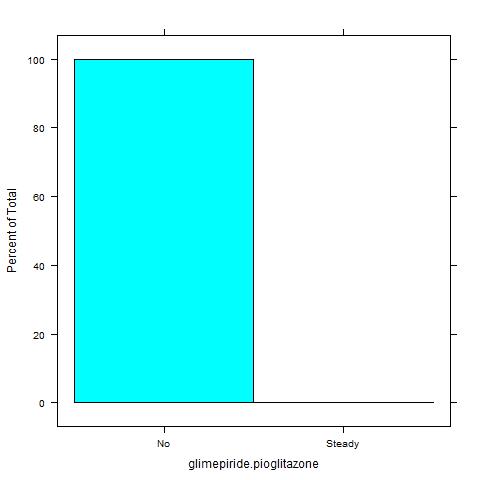
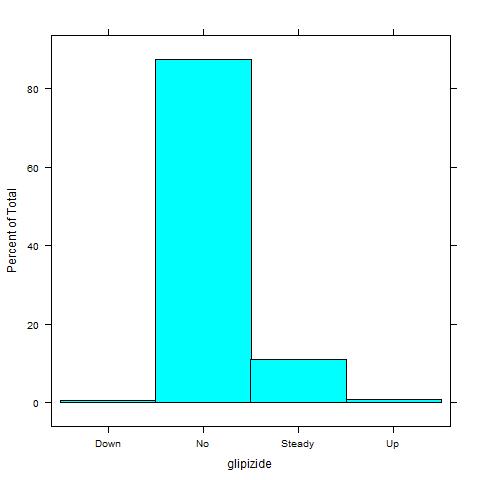
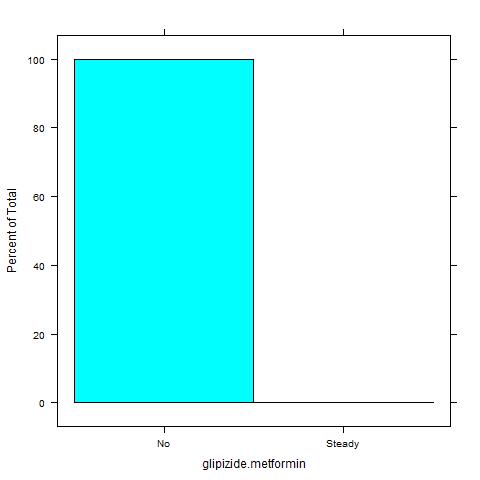
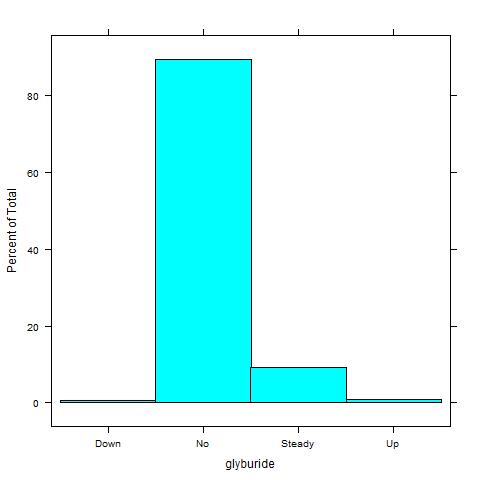
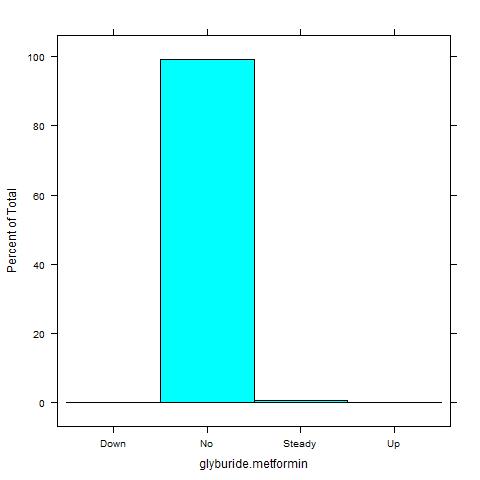
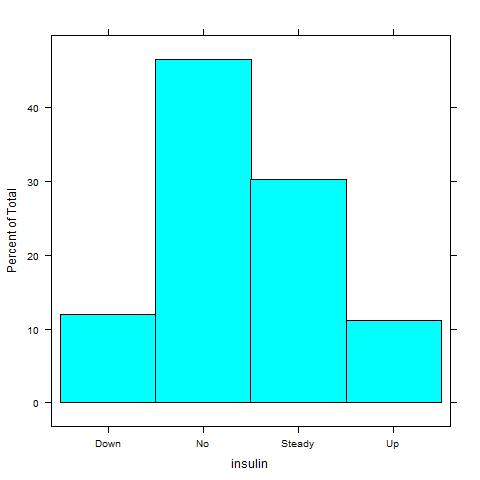
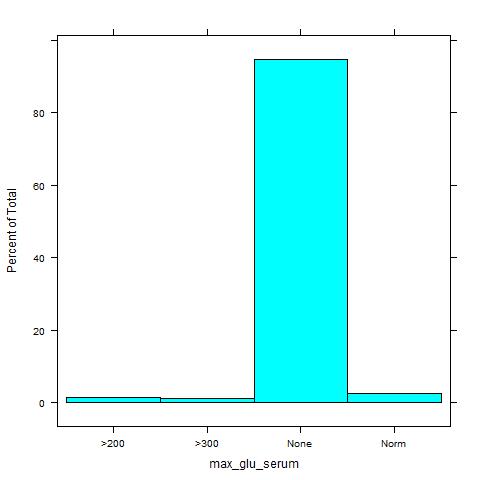
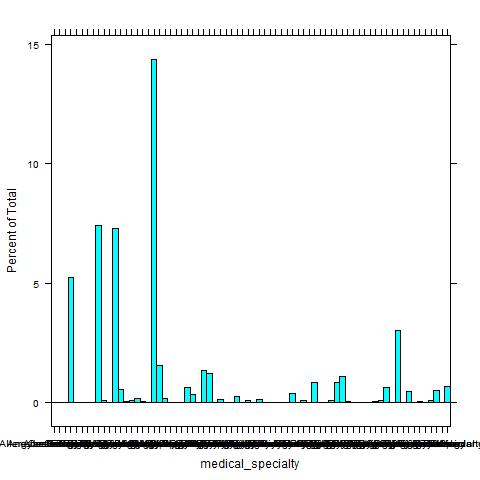
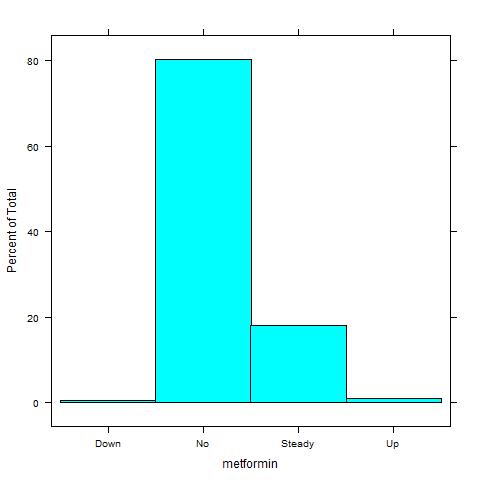
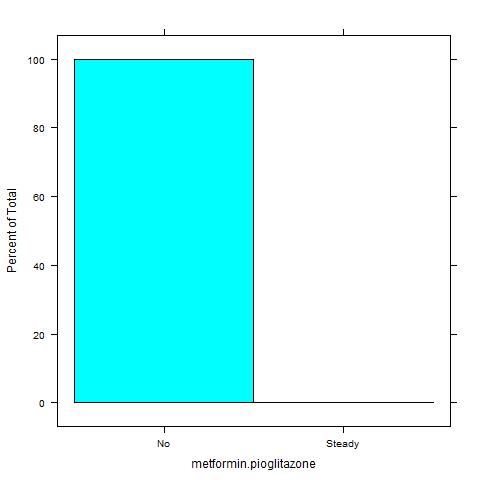
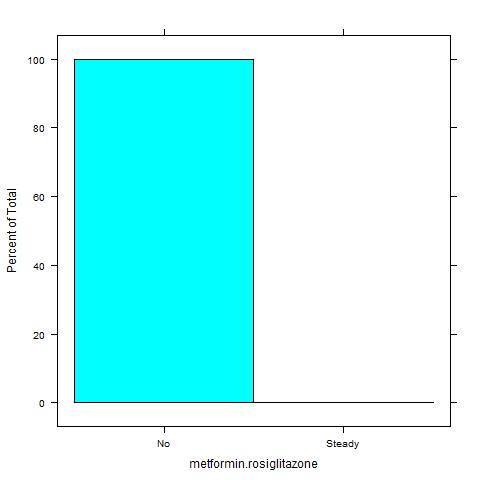
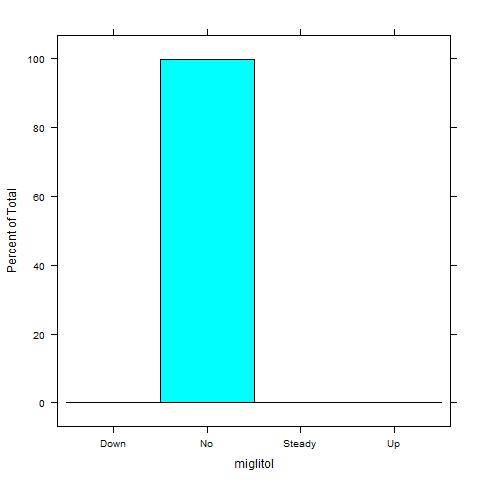
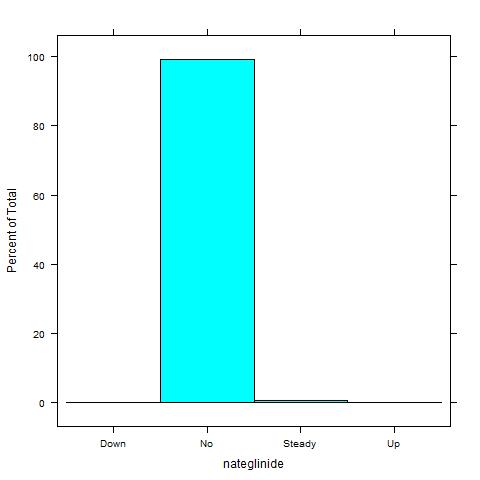
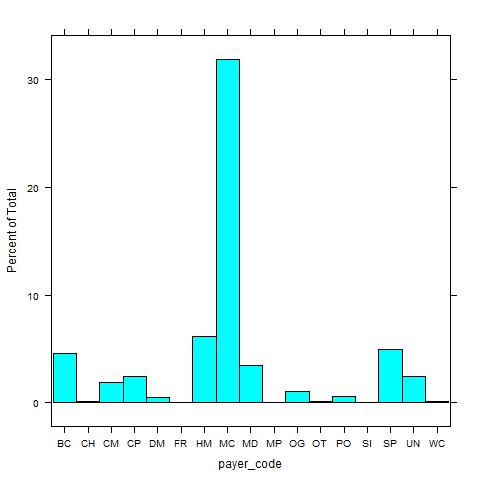
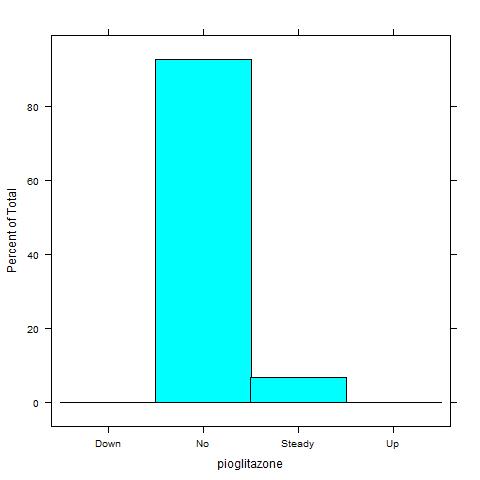
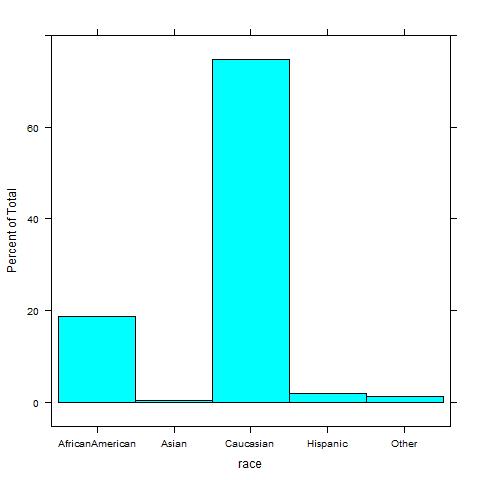
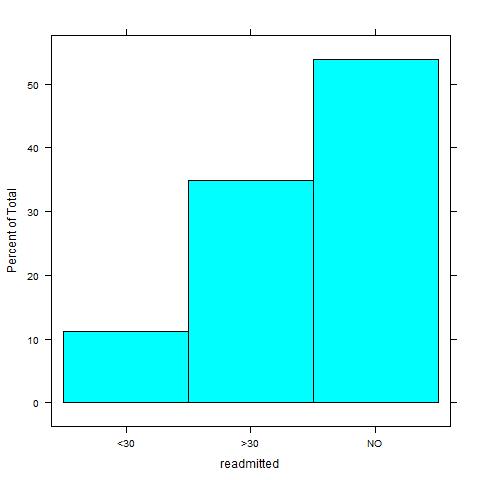
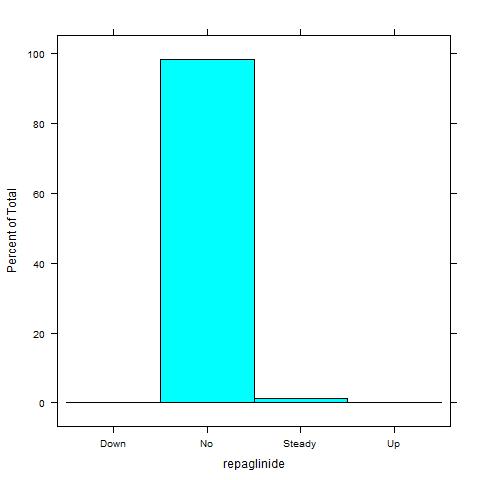
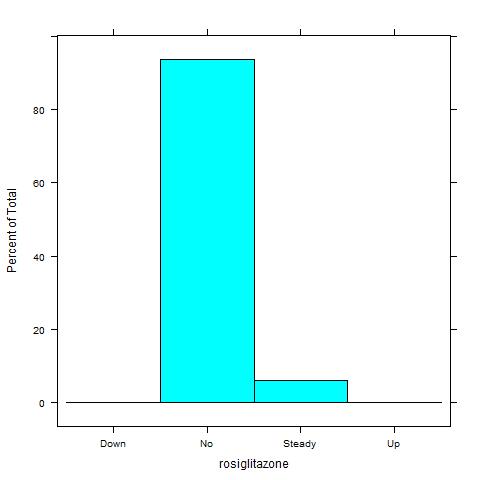
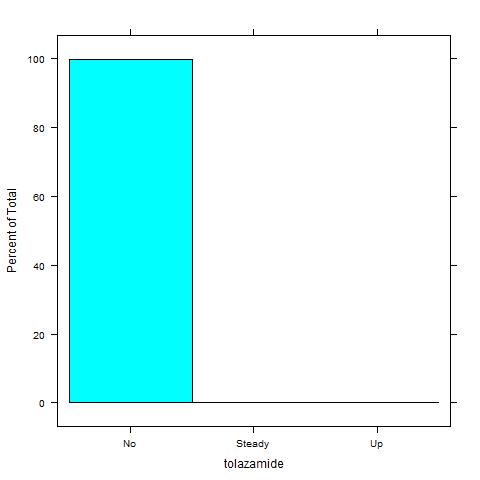
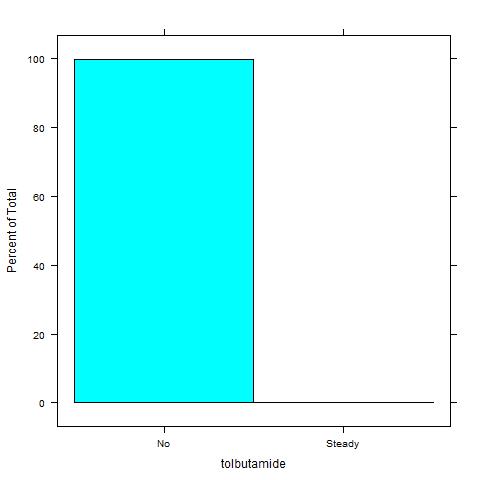
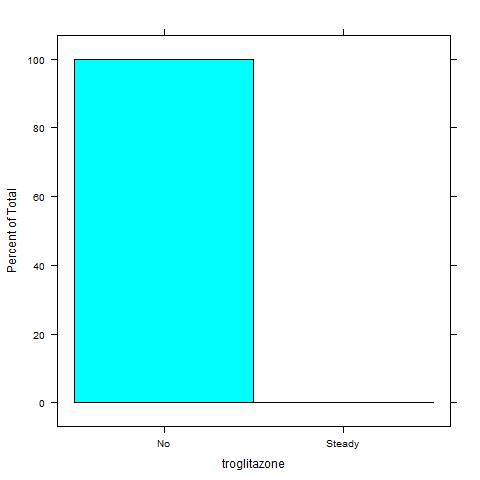
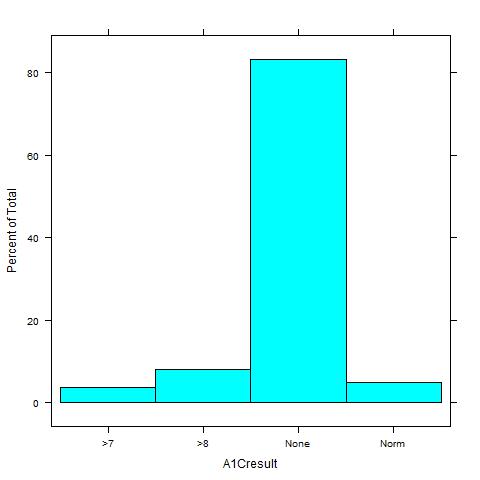
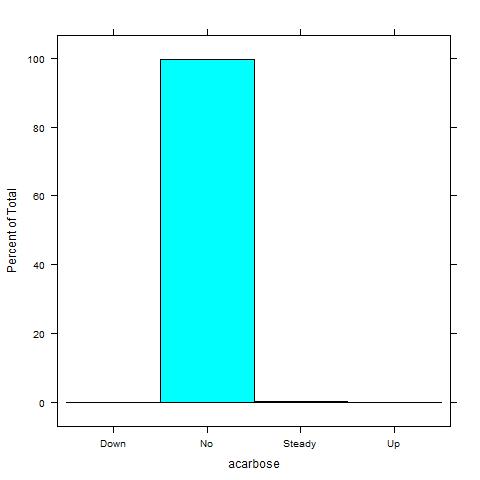
**Box – Plots:**

Box-plots tell us about outliers, we can clearly see outliers but in medical cases we cannot remove them because of extremities that can be occurred.



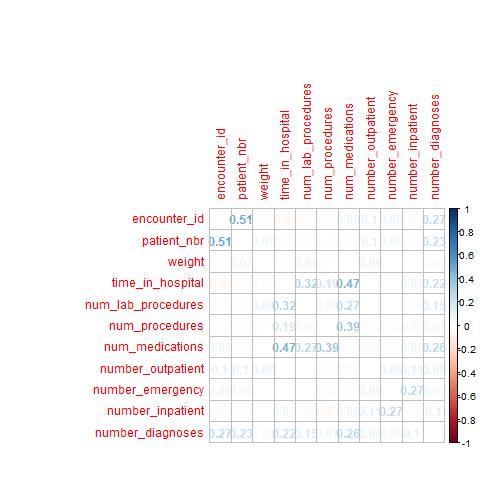
**Histograms:**

****

****

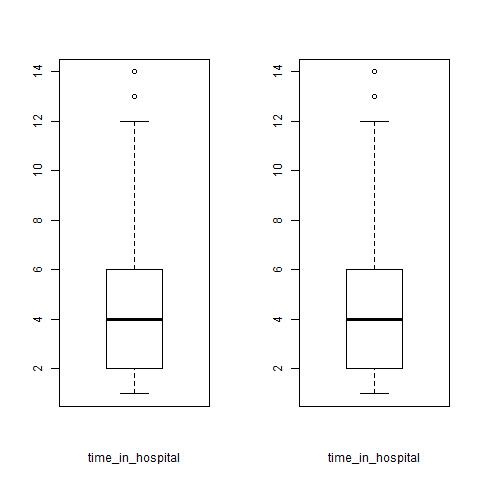
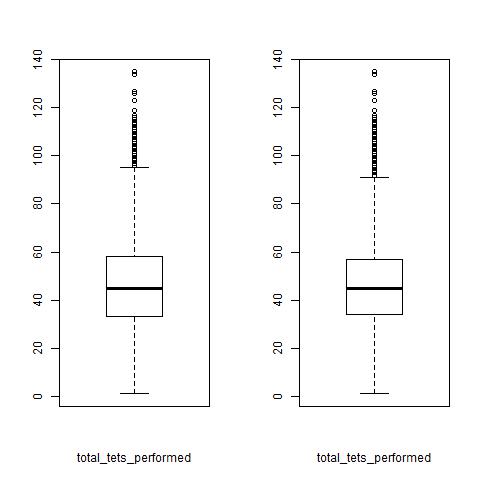
**Correlation Plot:**

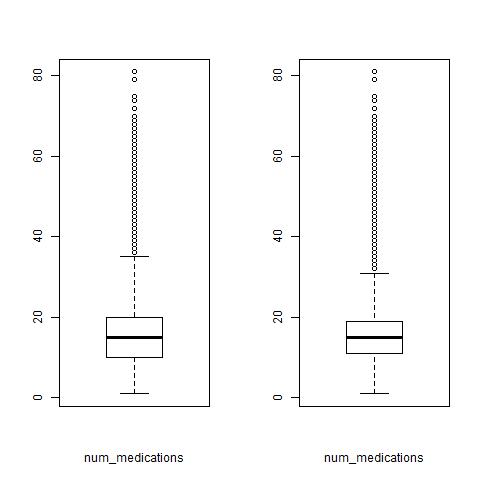
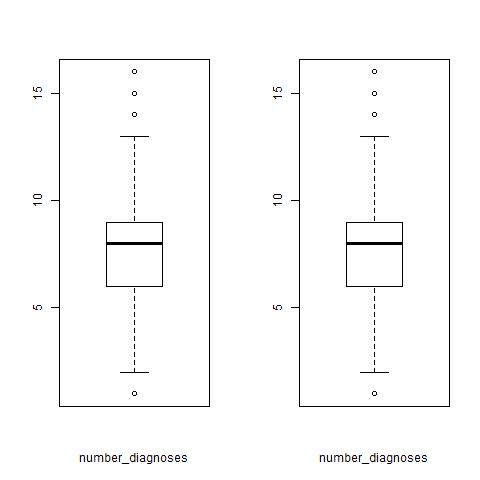
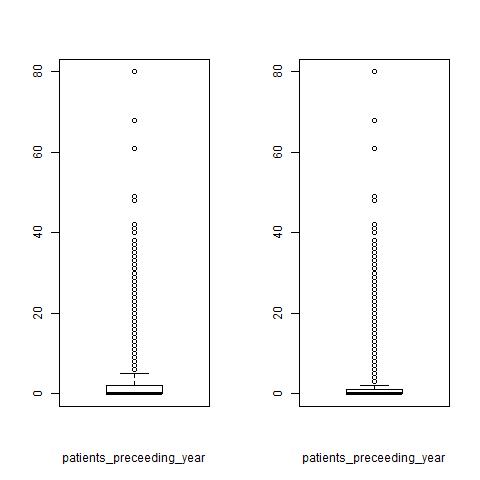
By correlation plot we can see that there is (no high correlation between the variables) low Multicollinearity.

****

**Box-plots: (Before and after)**

By applying central Imputation the numeric variables behaved in the following way.





By observing above box-plots, we know that there is not much change before and after the imputation.

Knn Imputation took a lot of time to run and raised error: “rep (…………)”.

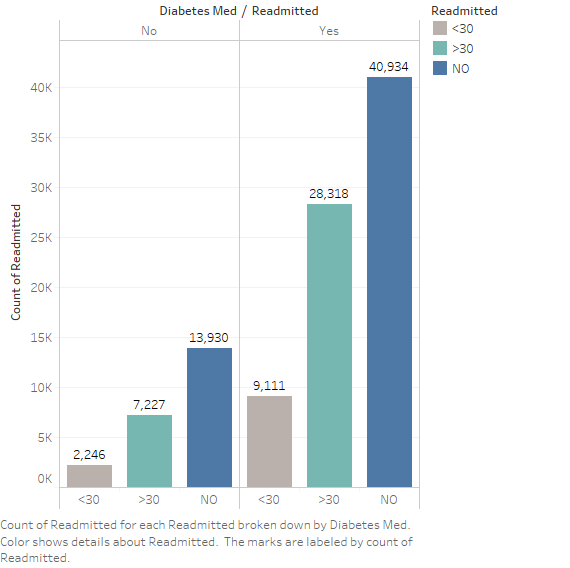
**Feature Generation:**

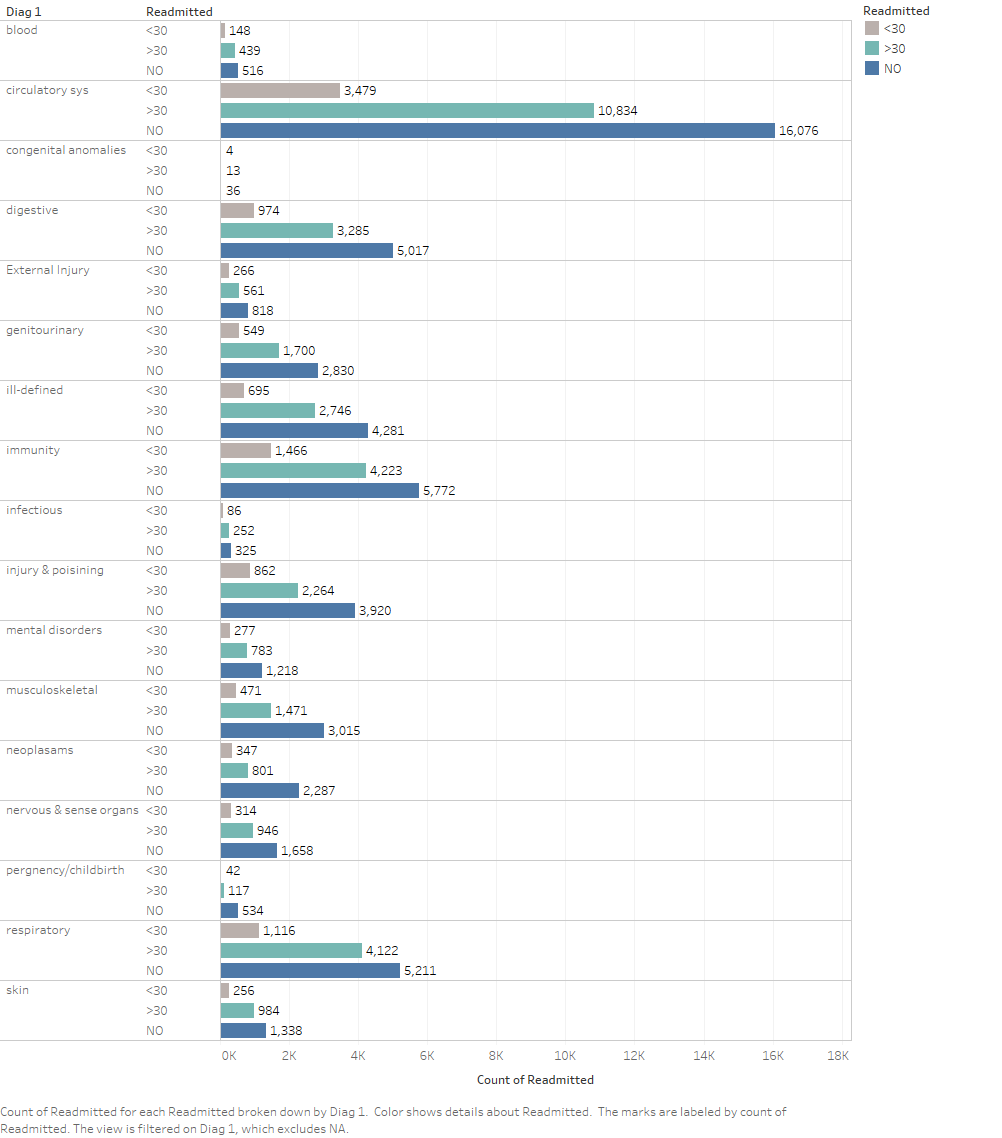
We have two new features which are derived by combining attributes.

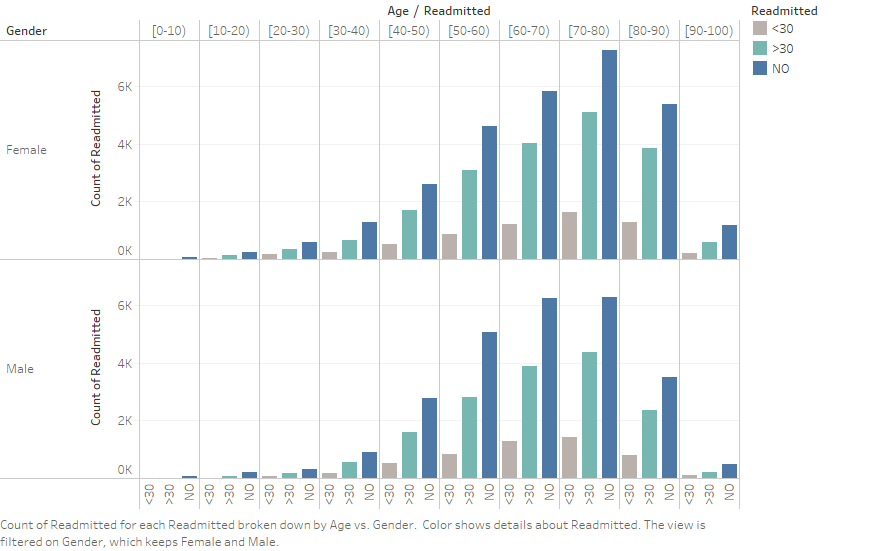
* **Total Tests Performed:**

Number of procedures and number of lab procedures of the patients

gives us total number of tests performed.

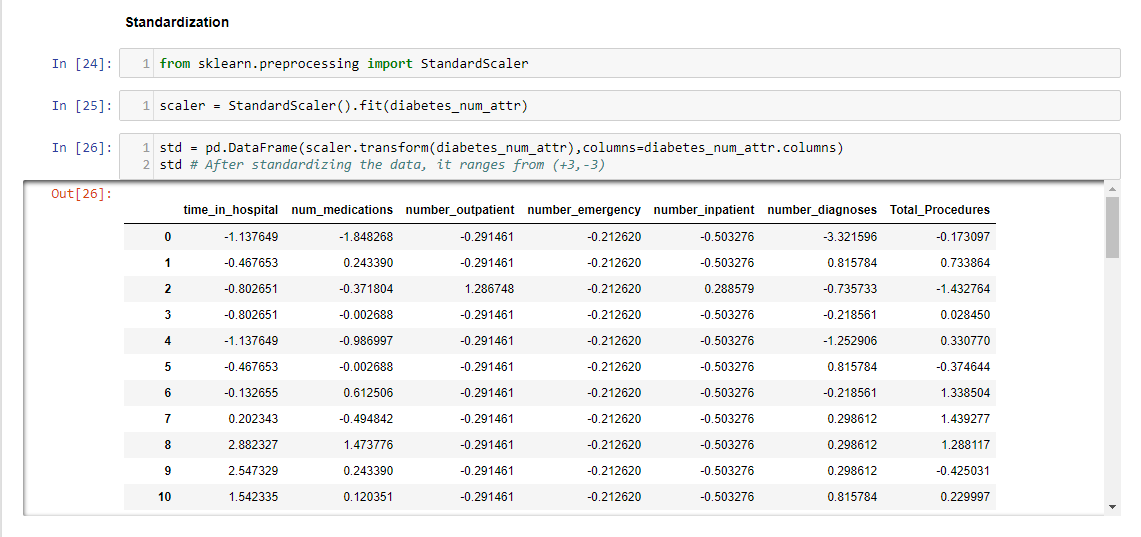






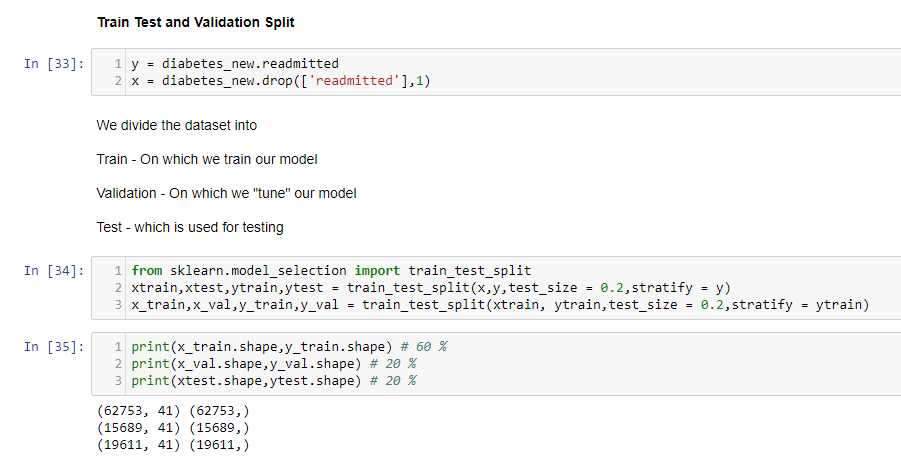
**Data Standardization:**

Standardization is done to make the attributes unitless. Results after standardization are as follows:



**Train-Test-Validation split:**

Train- 60%, Test- 20%, Validation- 20%. There is no thumb rule for this many people so this split and 70-10-20 is also done mostly.

****

**Model Building:**

**APPROACH -1:**

1. **Logistic Regression:**

Classification Report:

precision recall f1-score support

0 0.18 0.39 0.24 2213

1 0.43 0.27 0.33 6930

2 0.63 0.62 0.63 10468

avg / total 0.51 0.47 0.48 19611

Accuracy is 0.47325480597623781

1. **Decision Tree:**

Decision Tree classifier with criterion – ‘Gini Index’ (classification and Regression Tree) is used for modeling.

Classification Report:

precision recall f1-score support

0 0.20 0.26 0.23 2213

1 0.42 0.23 0.30 6930

2 0.61 0.75 0.67 10468

avg / total 0.50 0.51 0.49 19611

Accuracy is 0.514252205395

1. **SVM:**

Support Vector Machine – Linear is used. Kernels – radial, polynomial are over fitting the

model. These kernels are computationally expensive.

Classification Report:

precision recall f1-score support

0 0.26 0.10 0.15 2213

1 0.56 0.02 0.03 6930

2 0.55 0.98 0.70 10468

avg / total 0.52 0.54 0.40 19611

Accuracy is 0.53730049462

1. **Random Forest:**

This model is one of the benchmark models, it uses one of the ensemble techniques (Bagging).

Classification Report:

precision recall f1-score

0 0.35 0.03 0.06

1 0.47 0.39 0.43

2 0.61 0.79 0.69

avg / total 0.53 0.56 0.52

Accuracy is 0.564726878705

1. **XGBoost: (Extreme Gradient Boosting)**

Classification Report:

precision recall f1-score support

0 0.35 0.06 0.10 2213

1 0.52 0.36 0.42 6930

2 0.61 0.84 0.71 10468

avg / total 0.55 0.58 0.54 19611

Accuracy is 0.580898475345

1. **Majority Voting:**

Classification Report:

precision recall f1-score support

0 0.22 0.08 0.11 2213

1 0.50 0.37 0.42 6930

2 0.61 0.80 0.69 10468

avg / total 0.53 0.57 0.53 19611

Accuracy is 0.567028708378

1. **Stacking:**

precision recall f1-score support

0 0.03 0.31 0.05 211

1 0.40 0.49 0.44 5662

2 0.80 0.61 0.69 13738

avg / total 0.67 0.57 0.61 19611

Accuracy: 0.569272347152

**APPROACH – 2:**

In this approach, we consider only the important features, and do the model building. Results are as follows

Logistic:

precision recall f1-score support

0 0.18 0.40 0.24 2213

1 0.42 0.27 0.33 6930

2 0.63 0.61 0.62 10468

avg / total 0.51 0.47 0.48 19611

Accuracy: 0.46744174188

Xgboost:

precision recall f1-score support

0 0.04 0.37 0.07 242

1 0.39 0.51 0.44 5227

2 0.83 0.61 0.71 14142

avg / total 0.70 0.58 0.63 19611

Accuracy: 0.583805007394

**APPROACH – 3:**

In this approach, class imbalance is handled in a different way (without using SMOTE).

* Target has 3 Classes – NO, >30, <30
* Combined >30, <30 into one class and named it YES
* Built two separate models for each of them. (One with target – Y/N, other with target - >30/<30)

**Target – Y/N:**

Models Built on them and their results are as follows:

Logistic Regression:

precision recall f1-score support

1 0.63 0.39 0.48 7230

2 0.61 0.80 0.69 8459

avg / total 0.62 0.61 0.59 15689

Accuracy: 0.61233985595

Random Forest:

precision recall f1-score support

1 0.61 0.55 0.58 7230

2 0.64 0.70 0.67 8459

avg / total 0.63 0.63 0.63 15689

Accuracy is 0.628529542992

**Target - >30/<30:**

Logistic Regression:

precision recall f1-score support

0 0.01 0.45 0.02 62

1 1.00 0.76 0.86 9081

avg / total 0.99 0.75 0.85 9143

accuracy is 0.754675708192

Random Forest:

precision recall f1-score support

0 0.04 0.56 0.08 174

1 0.99 0.76 0.86 8969

avg / total 0.97 0.76 0.85 9143

accuracy is 0.757738160341

The above are the three approaches we consider.

Error Metric

Recall is the error metric.

Recall = True Positive / Total actual Positive

Model should correctly classify patient’s readmission rate, it costs a patient’s life if medication is delayed and also hospital’s reputation.