Data Analysis Report:

Data obtained from:
https://data.world/uci/diabetes-130-us-hospitals-for-years-1999-2008

The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. John Clore (jclore '@' vcu.edu), Krzysztof J. Cios (kcios '@' vcu.edu), Jon DeShazo (jpdeshazo '@' vcu.edu), and Beata Strack (strackb '@' vcu.edu). This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO).

Parameters

Race - Caucasian, African American, Asian, Hispanic, Other.

Gender - Female, Male, Unknown

Age - [0-10), [10-20), [20-30), [30-40), [40-50), [50-60), [60-70), [70-80), [80-90), [90-100)

Weight - pounds/lbs

Admission Type - Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available

Discharge Disposition - Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available

Admission Source - Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital

Time in Hospital - Integer number of days between admission and discharge

Payer Code - Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay

Medical Specialty - Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon

Number of Lab Procedures - Number of lab tests performed during the encounter

Number of Procedures - Number of procedures (other than lab tests) performed during the encounter

Number of Medications - Number of distinct generic names administered during the encounter

Number of Outpatient Visits - Number of outpatient visits of the patient in the year preceding the encounter

Number of Emergency Visits - Number of emergency visits of the patient in the year preceding the encounter

Number of Inpatient Visits - Number of inpatient visits of the patient in the year preceding the encounter

Primary Diagnosis - The primary diagnosis (coded as first three digits of ICD9); 848 distinct values

Secondary Diagnosis - Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values

Additional Secondary Diagnosis - Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values

Number of Diagnoses - Number of diagnoses entered to the system

Max Glucose Serum - Indicates the range of the result or if the test was not taken. Values: >200, >300, normal, and none if not measured

A1Cresult - Indicates the range of the result or if the test was 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.

Drugs - there are 25 drug parameters. Each drug is assigned the following 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

Change - Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change

Diabetes Medication - Indicates if there was any diabetic medication prescribed. Values: yes and no

Readmitted - Days to inpatient readmission. Values: <30 if the 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.

Missing Values

Discharge disposition, medical specialty, payer code will not be used in the data. A1Cresult, and Max Glucose Serum, and diagnosis types were also not used.

After analyzing through testing, drug types were also disregarded. This was due to the fact that the vast majority of the drug values were assigned as no, meaning they were not prescribed.

Given the specific context of the project, which is to project outcomes for patients based on race and gender, and to determine the quality of patient care and if there was any bias or mistreatment of a patient, we cannot use methods like data replacement for missing values in a record. Given that our source has over 100,000 records, we will only choose records that do not have missing values. We know that the only parameters that do have missing values are race, gender, and weight. This means we can simply filter out these records. We may look into selecting records based on race and gender that match the distribution of these categories across the population of the United States based on a census.

Transformations

Due to the large amount of categorical data, we will be using label encoding to transform these parameters into quantitative values.

No dimensionality reduction will be taking place. This decision was made in order to preserve the features of the data. Since we want to project outcomes for patients based on race and gender, dimensionality reduction has the possibility of removing the context of these features when translating them into different dimensional space.

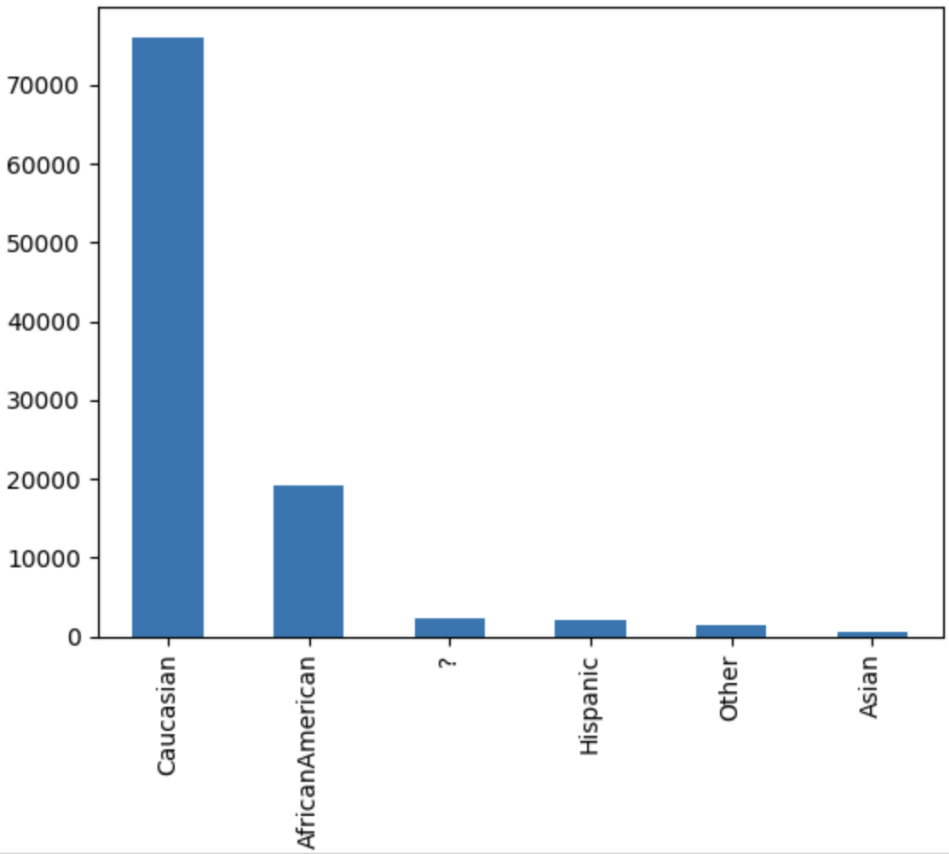
Separation

We will use k fold cross validation, meaning our folds will be comprised of training data and validation data, while we set aside testing data.

Visualization

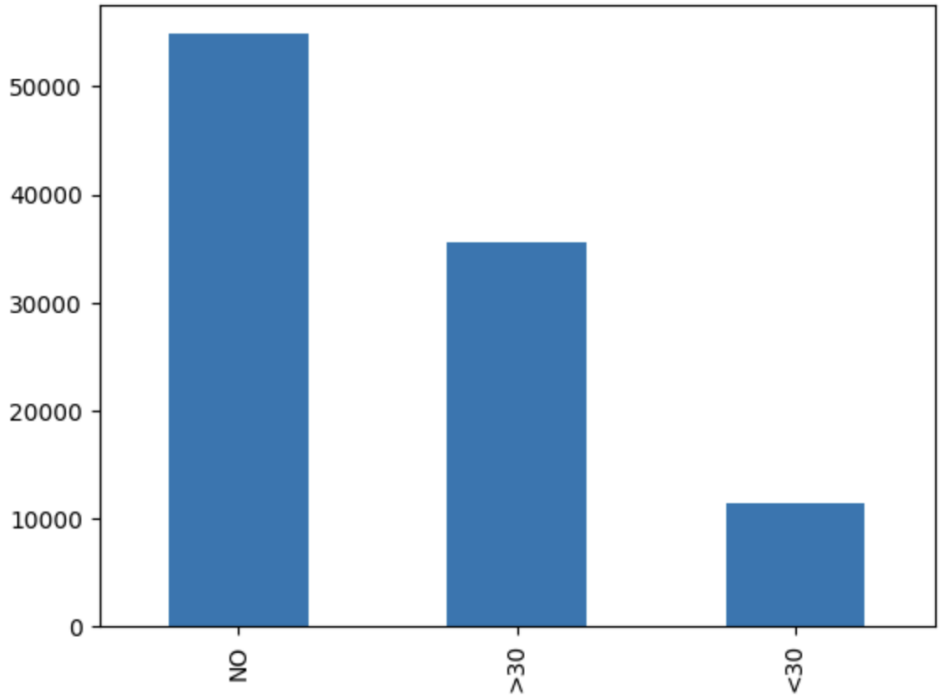
As there are a large number of parameters, race, readmissions, change, inpatient visits, and number of procedures were chosen for preliminary data visualization.

```
<Axes: >
```


70000
60000
50000
40000
30000
20000
10000
0
Caucasian
AfricanAmerican
?
Hispanic
Other
Asian

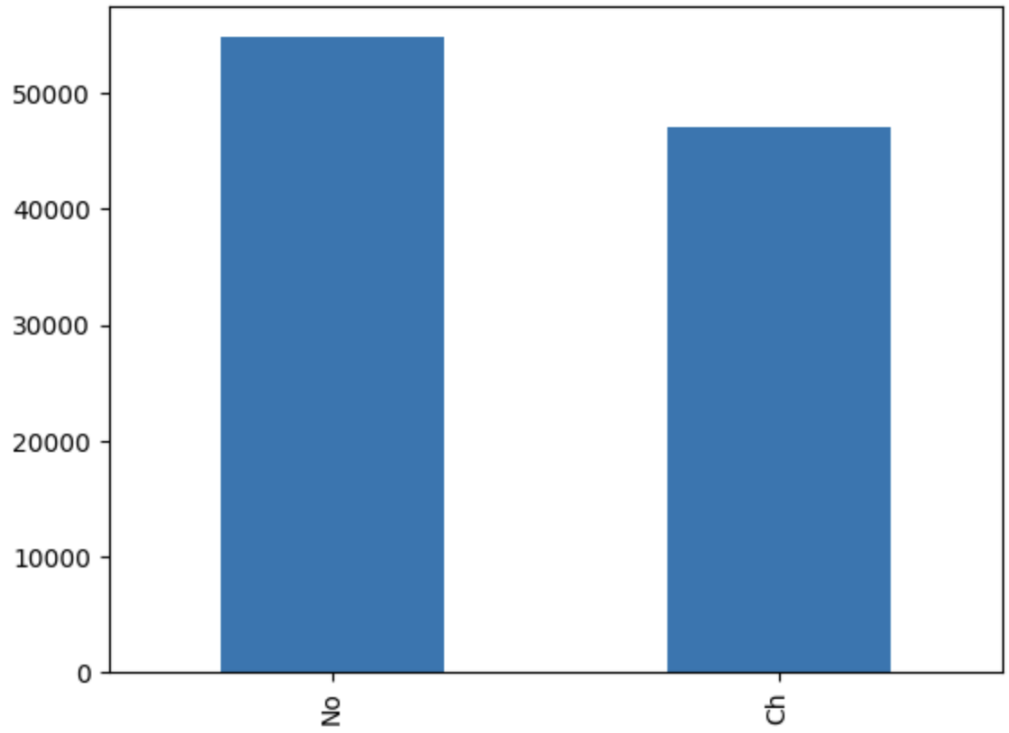
```
df1.readmitted.value_counts().plot(kind='bar')
```

```
<Axes: >
```



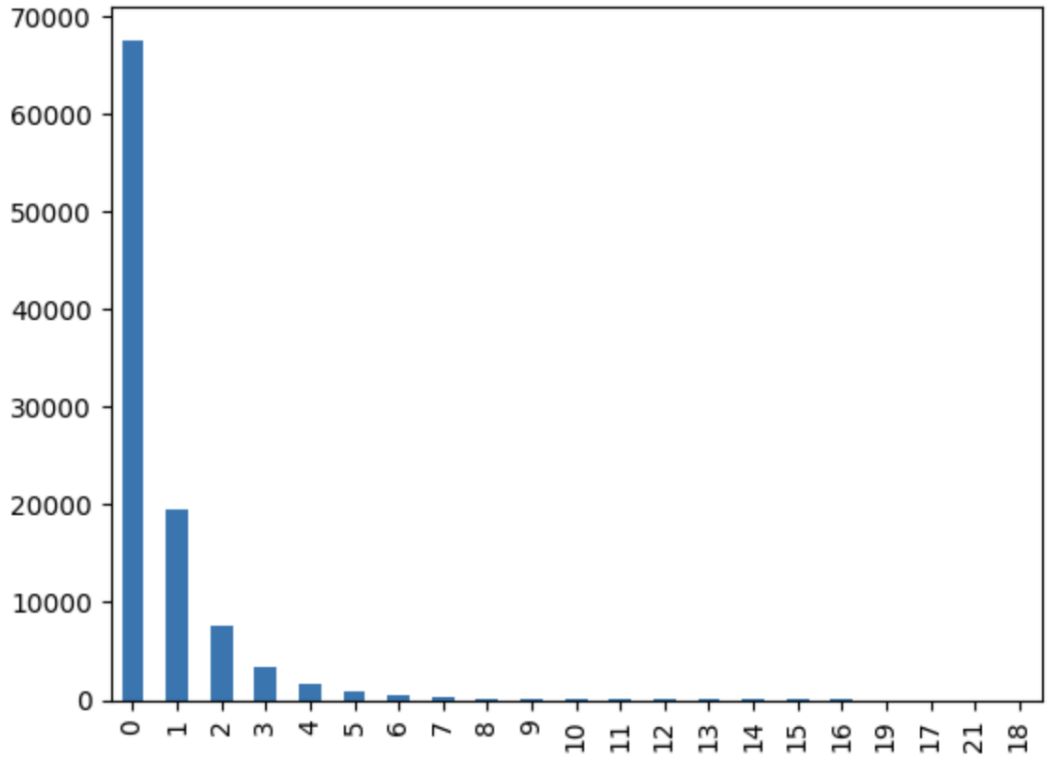
```
df1.change.value_counts().plot(kind='bar')
```

```
<Axes: >
```



```
df1.number_inpatient.value_counts().plot(kind='bar')
```

```
<Axes: >
```



```
df1.num_procedures.value_counts().plot(kind='bar')
```

```
<Axes: >
```