Project Progress

Dataset Introduction:

Dataset Name: Diabetes 130-US hospitals for years 1999-2008 Data Set

Dataset Source: http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

Topic: Healthcare

- Dataset is attached alongside this document.

Variable Summary:

Feature name	Type	IV/DV	Description and values	% missing
Encounter ID	Primary Key	IV	Unique identifier of an encounter	0%
Patient number	Primary Key	IV	Unique identifier of a patient	0%
Race	Nominal	IV	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	IV	Values: male, female, and unknown/invalid	0%
Age	Interval	IV	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)	0%
Weight	Ratio	IV	Weight in pounds.	97%
Admission type	Nominal	IV	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	IV	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	IV	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	**Ratio**	**DV**	**Integer number of days between admission and discharge**	**0%**
Payer code	Nominal	IV	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	IV	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%

Feature name	Type	IV/DV	Description and values	% missing
Number of lab procedures	Ratio	IV	Number of lab tests performed during the encounter	0%
Number of procedures	Ratio	IV	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Ratio	IV	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Ratio	IV	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Ratio	IV	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Ratio	IV	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	IV	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	IV	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	IV	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Ratio	IV	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	IV	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	IV	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.	0%
Change of medications	**Nominal**	**DV**	**Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”**	**0%**
Diabetes medications	Nominal	IV	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Nominal	IV	For the generic names. 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	0%
Readmitted	Nominal	IV	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.	0%

Project Idea No.1:

Goal: (**Regression**) Predict Time in hospital variable which is an integer number of days between admission and discharge.

Business view: If we can predict time in the hospital, we can optimize hospital room usage, so we will have available rooms for all incoming patients.

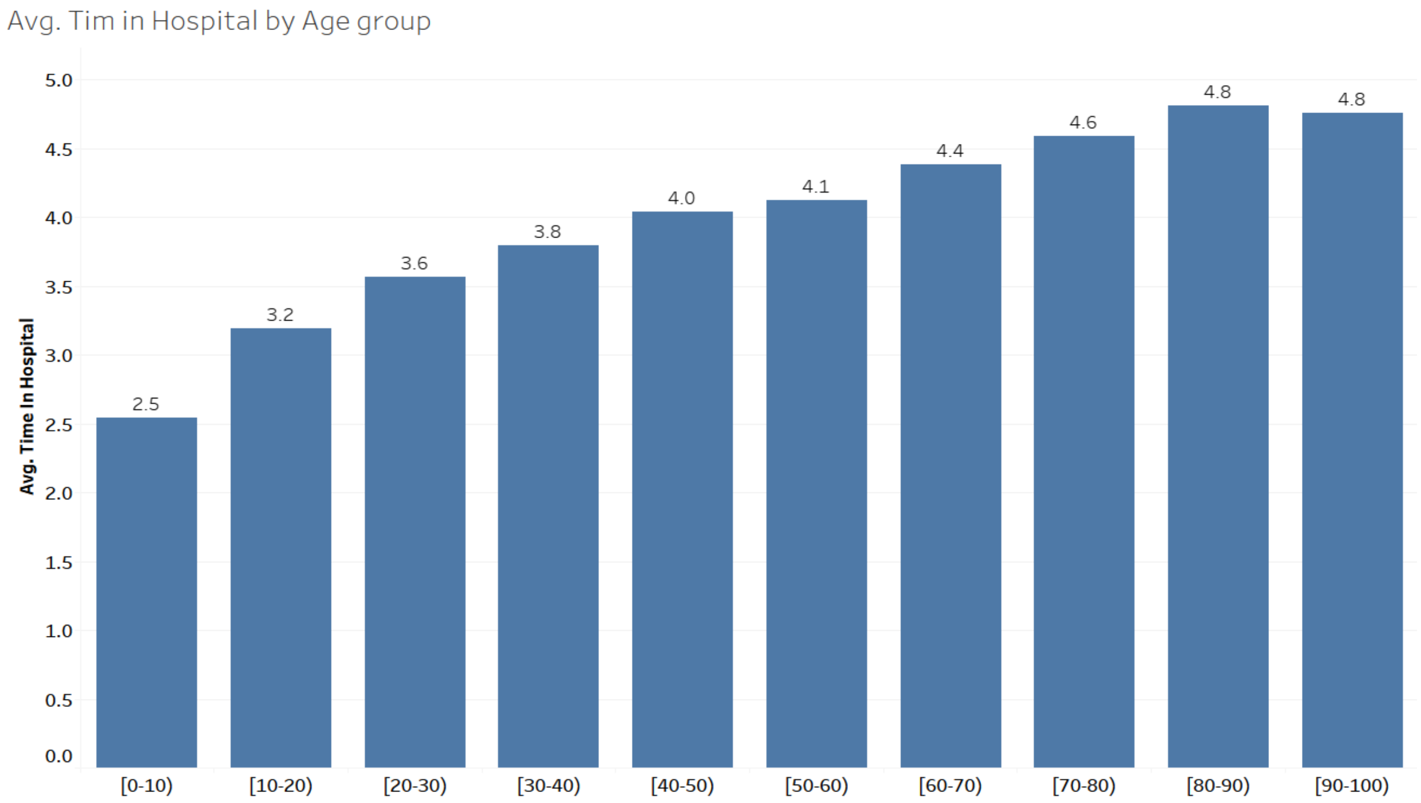
Project Idea No.2:

Goal: (**Classification**) Predict Change of medications variable which Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “**change**” and “**no change**”.

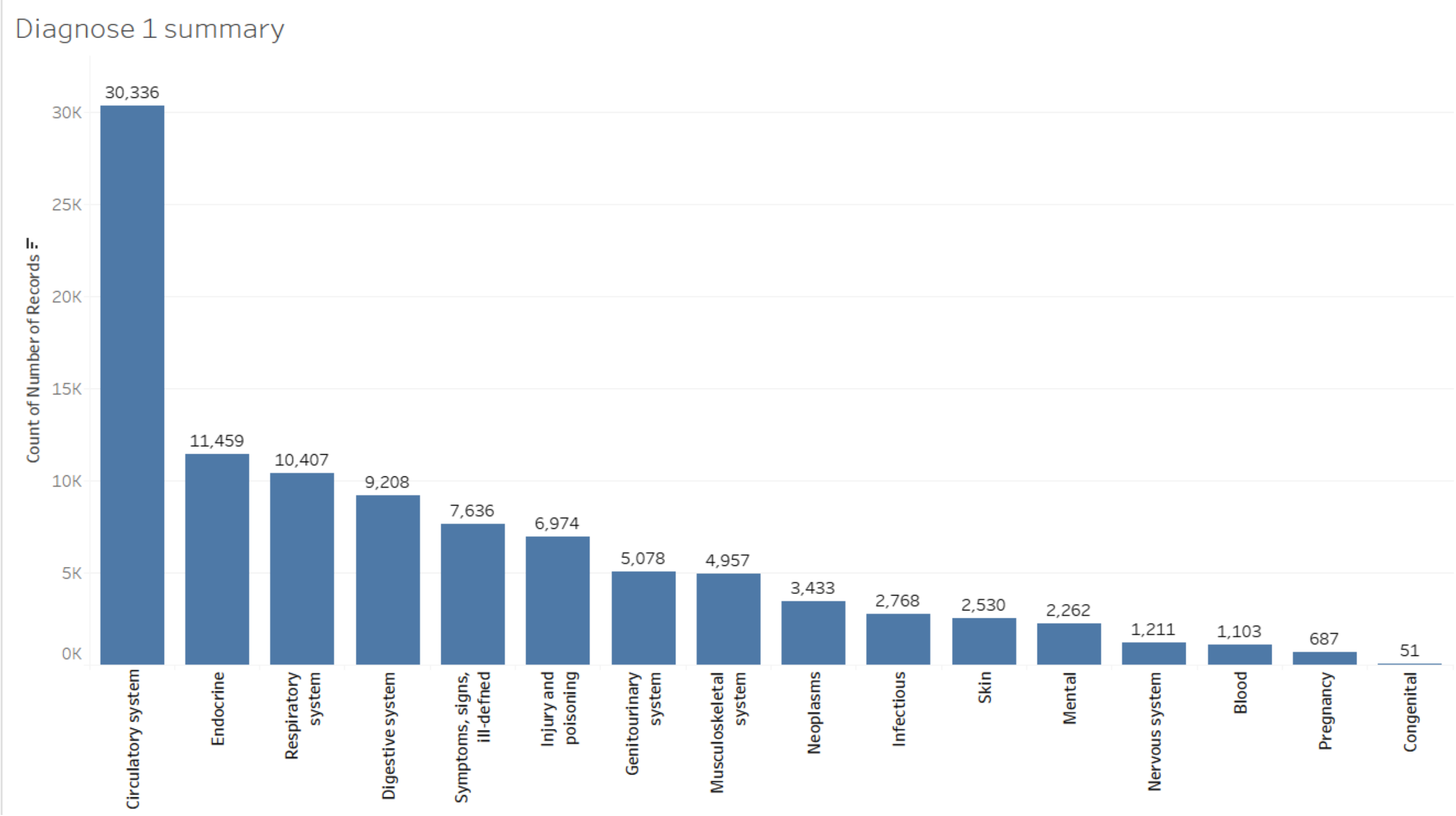
Business view: Hospitals can create an online predictor for patients, and after inserting their data, if they are classified into “**Change**” medication, they need to go to the hospital to get checked for further diagnosis and prescription.

Descriptive model (on numerical variables):

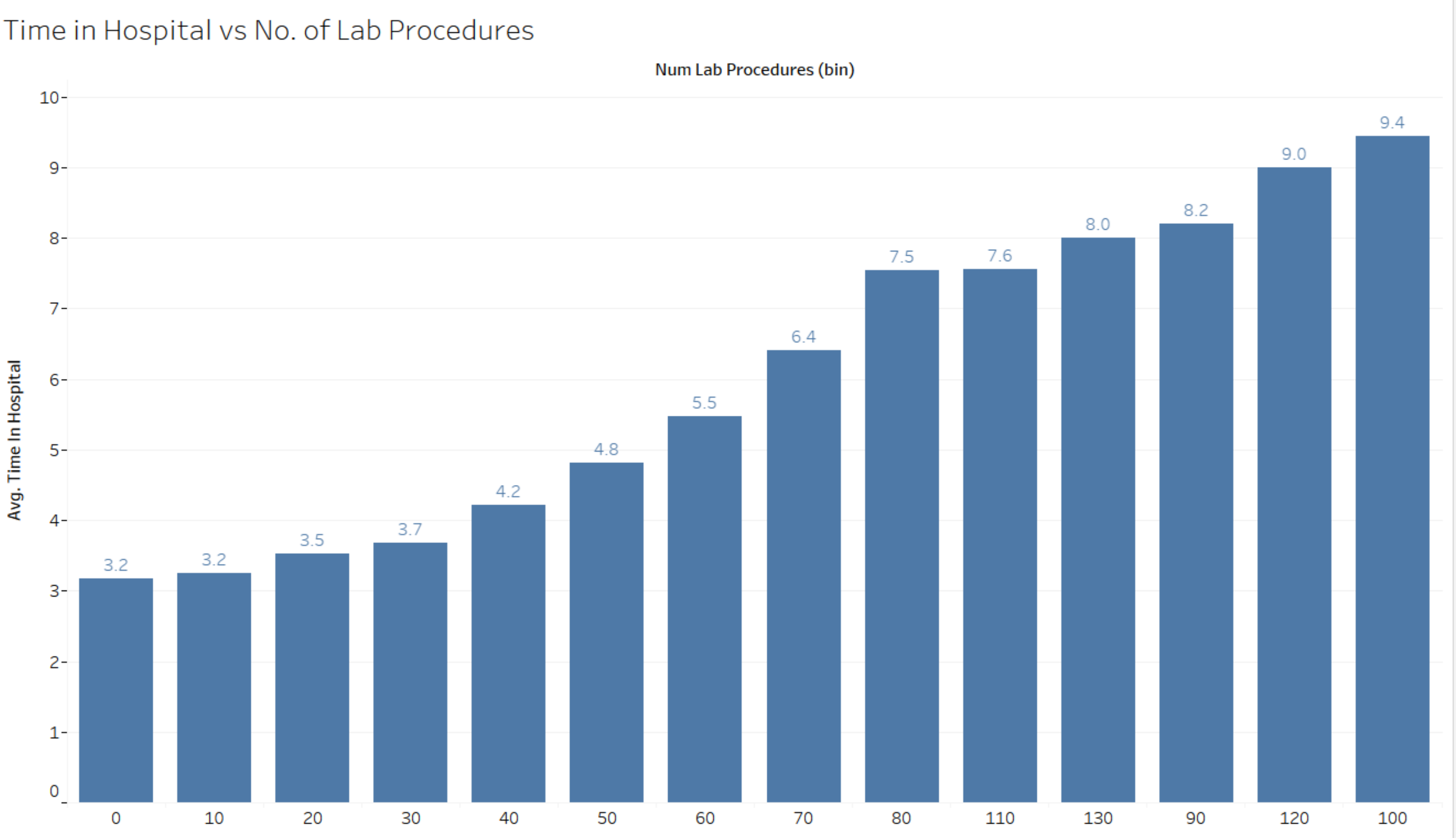
Column Name	Min	Max	Mean	Std. Deviation	Variance	No. Missing
admission_source_id	1	25	5.7544	4.0641	16.5168	0
admission_type_id	1	8	2.0240	1.4454	2.0892	0
discharge_disposition_id	1	28	3.7156	5.2802	27.8801	0
encounter_id	12522	4.44E+08	1.65E+08	1.03E+08	1.05E+16	0
num_lab_procedures	1	132	43.0956	19.6744	387.0805	0
num_medications	1	81	16.0218	8.1276	66.0573	0
num_procedures	0	6	1.3397	1.7058	2.9098	0
number_diagnoses	1	16	7.4226	1.9336	3.7388	0
number_emergency	0	76	0.1978	0.9305	0.8658	0
number_inpatient	0	21	0.6356	1.2629	1.5948	0
number_outpatient	0	42	0.3694	1.2673	1.6060	0
patient_nbr	135	1.90E+08	5.43E+07	3.87E+07	1.50E+15	0
time_in_hospital	**1**	**14**	**4.3960**	**2.9851**	**8.9109**	**0**



Age vs Time in Hospital



Diagnosis Distribution



Number of Lab procedures vs Time in Hospital

Predictive model (For Project idea no.1):

We performed a preliminary Linear Regression. Our Target variable is the "**Time_in_hospital**". We partitioned data to Training set and Testing set with ratio of 80:20. The performance of the model is observable through the following tables:

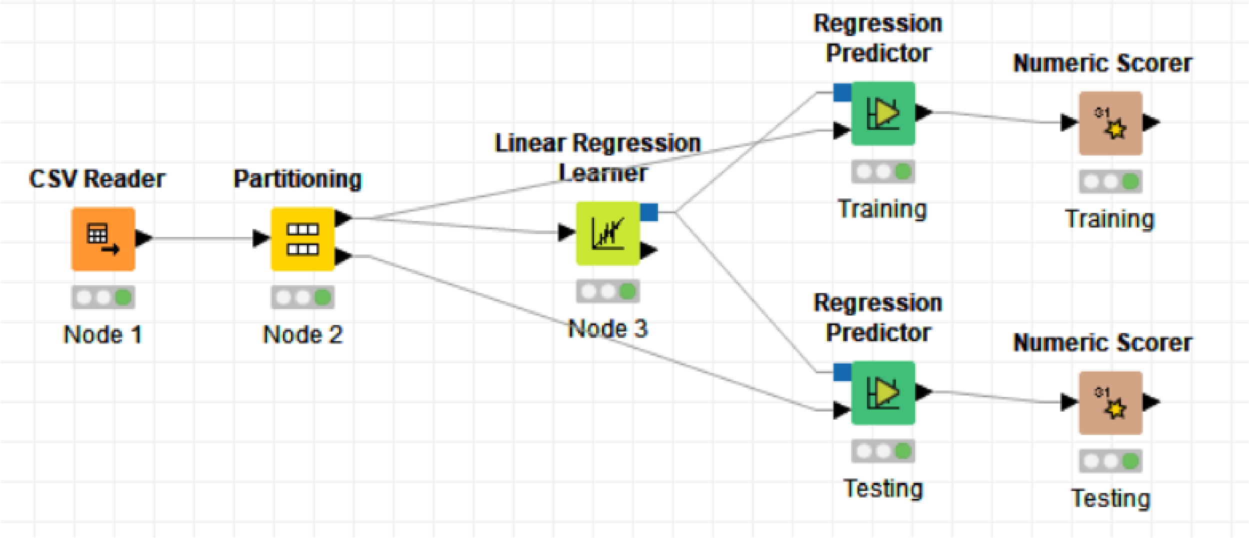
Variable	Coeff.	Std. Error	t-value	P>|t|
age	1.5759	0.0552	28.5739	0.0000
num_lab_procedures	3.9408	0.0628	62.7688	0.0000
num_procedures	0.5209	0.0358	14.5334	0.0000
num_medications	10.6666	0.1220	87.4203	0.0000
number_emergency	-5.0408	0.7079	-7.1209	0.0000
number_diagnoses	1.4846	0.0777	19.1110	0.0000
change	0.0371	0.0263	1.4071	0.1594
No_metformin	3.7988E+13	2.5127E+12	15.1182	0.0000
Steady_metformin	3.7988E+13	2.5127E+12	15.1182	0.0000

Variable	Coeff.	Std. Error	t-value	P>|t|
Up_metformin	3.7988E+13	2.5127E+12	15.1182	0.0000
Down_metformin	3.7988E+13	2.5127E+12	15.1182	0.0000
No_insulin	-1.0558E+13	1.3968E+12	-7.5586	0.0000
Up_insulin	-1.0558E+13	1.3968E+12	-7.5586	0.0000
Steady_insulin	-1.0558E+13	1.3968E+12	-7.5586	0.0000
Down_insulin	-1.0558E+13	1.3968E+12	-7.5586	0.0000
NO_readmitted	-2.9530E+13	1.8959E+12	-15.5754	0.0000
>30_readmitted	-2.9530E+13	1.8959E+12	-15.5754	0.0000
<30_readmitted	-2.9530E+13	1.8959E+12	-15.5754	0.0000
pregnancy_diag	0.3155	0.1093	2.8850	0.0039
Endocrine_diag	-0.0477	0.0211	-2.2604	0.0238
Infectious_diag	0.4657	0.0380	12.2470	0.0000
NEoplasms_diag	0.5402	0.0391	13.8163	0.0000
circulatory_diag	-0.4247	0.0236	-17.9718	0.0000
Respiratory_diag	0.2168	0.0243	8.9334	0.0000
Injury_diag	0.1711	0.0324	5.2726	0.0000
Skin_diag	1.0301	0.0356	28.9614	0.0000
musculoskeletal_diag	-0.1809	0.0355	-5.0956	0.0000
Digestive_diag	0.1771	0.0284	6.2395	0.0000
Congential_diag	-0.0465	0.1813	-0.2562	0.7978
Genitourinary_diag	0.0903	0.0258	3.5065	0.0005
symptoms_diag	-0.3729	0.0261	-14.2646	0.0000
Mental_diag	0.8314	0.0378	21.9743	0.0000
Nervous_diag	0.3059	0.0456	6.7071	0.0000
V_diag	0.6371	0.0364	17.4813	0.0000
E_diag	-0.5373	0.0690	-7.7905	0.0000
blood_diag	-0.1307	0.0374	-3.4901	0.0005
Intercept	2.1006E+12	4.3738E+12	0.4803	0.6310

Model performance:

Score Name	Training set	Test Set
R_Square	**0.297**	**0.292**
Mean Absolute Error	1.893	1.897
Mean Squared Error	6.274	6.275
Root Mean Squared Error	2.505	2.505
Mean Signed Difference	-0.067	-0.077
Mean absolute percentage Error	0.652	0.651

Workflow screenshot:



Interpreting Linear Regression Model:

The R_square of our model is around 0.297, which means that our model can predict around 30% of variation in our data, which is a good place to start but not enough. We believe we can improve that.

Age is one of the significant independent variables in predicting Time_in_Hospital. The Coefficient of Age is equal to 1.5759, which means if a patients Age, increases by one unit, the time in hospital also increases around 1.5759 days which make sense as older patients need more time for recovery.

Another significant IV is num_lab_procedures which indicates the number of lab tests patient needs to perform in their stay. As num_lab_procedures increase by one unit, the time in hospital increases by 3.9408 days. Again this seems to be right, as a patient who needs more lab test done should stay longer in hospital for their test result, analysis and further instruction.

ethical ramifications of data and analysis:

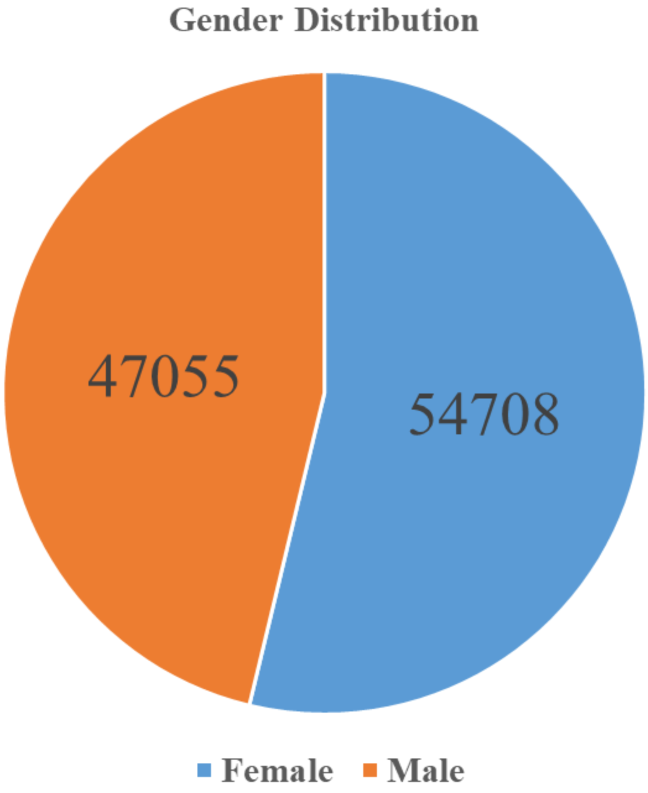
The Health Facts data we used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500. As our analytics is going to be used in US, the data is covering the whole variance of distribution of patients in different geographical places.

According to our data, the distribution of race is very close to the actual race distribution in US. In the following table we have included the actual distribution and our data distribution of races for comparison. The available difference is due to the fact that our data is collected from 1999 – 2008, and actual distribution is from 2018. Moreover, these differences can be justified because of nature of races and their tolerance against special diseases meaning that some races may be more vulnerable to special disease.

Race levels	Count	Percentage	Actual race Percentage in US*
Null	2,273	2.23%	-
African-American	19,210	18.88%	12.00%
Asian	641	0.63%	6.00%
Caucasian (White)	76,099	74.78%	60.00%
Hispanic	2,037	2.00%	18.00%
Other	1,506	1.48%	4.00%
Grand Total	**101,766**	**100%**	**100%**

***** reference: https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

In addition to that, the gender in our data is almost balanced between male patients and female patients with 53.76% females and 46.24% male patients:


Gender Distribution
🟦 Female
🟧 Male