Exploring Healthcare Through the Lens of Data.

by Neha Nayak.

Abstract

In this project, we explored the potential of data science in improving public health outcomes through a comprehensive analysis of real-world healthcare data. By examining the *Diabetes 130-US Hospitals* dataset using IBM SPSS, we applied statistical and machine learning techniques to identify patterns, understand trends, and predict hospital readmission. This project incorporated data preprocessing, exploratory data analysis (EDA), feature selection, regression modelling and hypothesis testing. The final outcomes offer actionable insights into the dynamics of hospital utilization and the influence of demographics and medications on patient readmissions.

Introduction

The intersection of healthcare and data science presents a powerful opportunity to uncover meaningful insights that can inform policy, improve patient outcomes, and streamline resource allocation. In this project we leveraged the *Diabetes 130-US Hospitals* dataset to perform in-depth statistical analysis using SPSS. Our goal was to clean and analyse the dataset, uncover trends, and use predictive models to forecast healthcare needs.

Methodology

This project followed a systematic approach over a six-week timeline:

Week 1-2: Data Preprocessing

Week 3: Exploratory Data Analysis

Week 4: Logistic Regression I

Week 5: Logistic Regression II

Week 6: ANOVA Hypothesis Testing

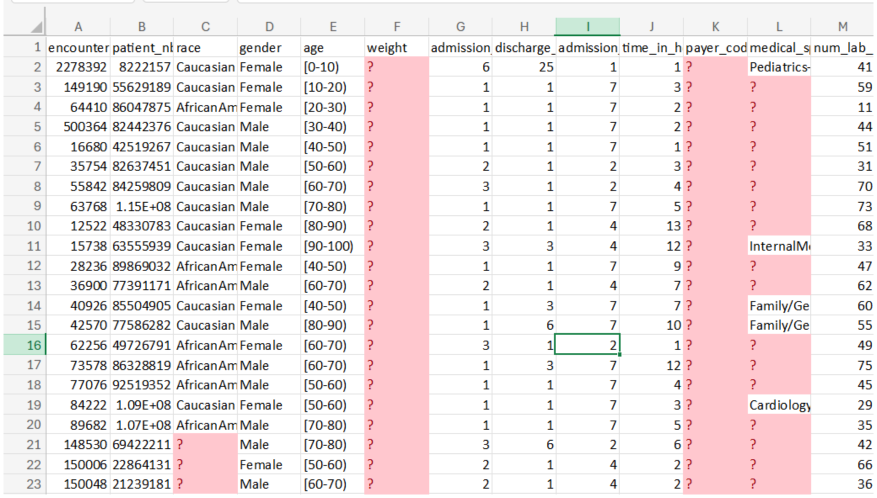
Claim Tested: Average hospital stay differs significantly across age groups.

- Ran descriptive stats and one-way ANOVA.
- Post-hoc Tukey test identified specific group differences.

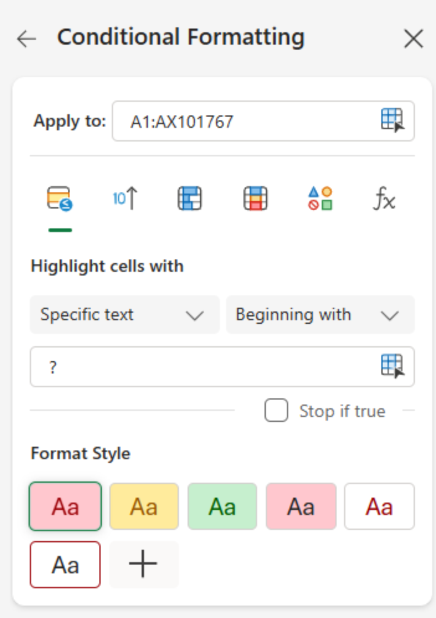
DATA PROCESSING

Dataset Used: Diabetes 130-US Hospitals
Tools: IBM SPSS
Steps:

STEP 1: Download the Diabetes csv file
STEP 2: Identifying and dealing with missing data



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	encounter	patient_nl	race	gender	age	weight	admission	discharge_	admission	time_in_h	payer_cod	medical_s	num_lab_
2	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	?	Pediatrics-	41
3	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	?	?	59
4	64410	86047875	AfricanAm	Female	[20-30)	?	1	1	7	2	?	?	11
5	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	?	?	44
6	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	?	?	51
7	35754	82637451	Caucasian	Male	[50-60)	?	2	1	2	3	?	?	31
8	55842	84259809	Caucasian	Male	[60-70)	?	3	1	2	4	?	?	70
9	63768	1.15E+08	Caucasian	Male	[70-80)	?	1	1	7	5	?	?	73
10	12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13	?	?	68
11	15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12	?	InternalM	33
12	28236	89869032	AfricanAm	Female	[40-50)	?	1	1	7	9	?	?	47
13	36900	77391171	AfricanAm	Male	[60-70)	?	2	1	4	7	?	?	62
14	40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7	?	Family/Ge	60
15	42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10	?	Family/Ge	55
16	62256	49726791	AfricanAm	Female	[60-70)	?	3	1	2	1	?	?	49
17	73578	86328819	AfricanAm	Male	[60-70)	?	1	3	7	12	?	?	75
18	77076	92519352	AfricanAm	Male	[50-60)	?	1	1	7	4	?	?	45
19	84222	1.09E+08	Caucasian	Female	[50-60)	?	1	1	7	3	?	Cardiology	29
20	89682	1.07E+08	AfricanAm	Male	[70-80)	?	1	1	7	5	?	?	35
21	148530	69422211	?	Male	[70-80)	?	3	6	2	6	?	?	42
22	150006	22864131	?	Female	[50-60)	?	2	1	4	2	?	?	66
23	150048	21239181	?	Male	[60-70)	?	2	1	4	2	?	?	36



Conditional Formatting

Apply to: A1:AX101767

Highlight cells with

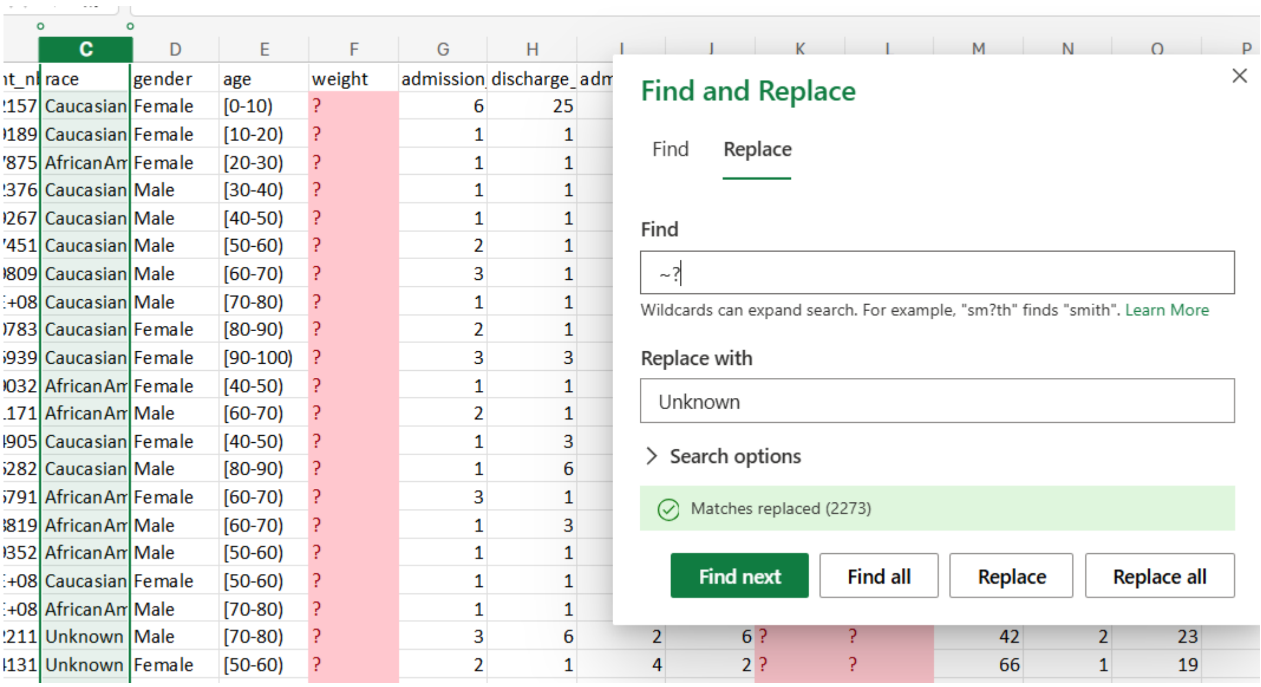
Specific text | Beginning with

?

☐ Stop if true

Format Style

- Missing values in Categorical Columns: *race*, *weight*, *payer_code*, *medical_specialty*, *diag_1*, *diag_2*, *diag_3*.



	C	D	E	F	G	H	I	J	K	L	M	N	O
it_nl	race	gender	age	weight	admission	discharge_	adm						
157	Caucasian	Female	[0-10)	?	6	25							
189	Caucasian	Female	[10-20)	?	1	1							
875	AfricanAm	Female	[20-30)	?	1	1							
376	Caucasian	Male	[30-40)	?	1	1							
267	Caucasian	Male	[40-50)	?	1	1							
451	Caucasian	Male	[50-60)	?	2	1							
809	Caucasian	Male	[60-70)	?	3	1							
+08	Caucasian	Male	[70-80)	?	1	1							
783	Caucasian	Female	[80-90)	?	2	1							
939	Caucasian	Female	[90-100)	?	3	3							
032	AfricanAm	Female	[40-50)	?	1	1							
171	AfricanAm	Male	[60-70)	?	2	1							
905	Caucasian	Female	[40-50)	?	1	3							
282	Caucasian	Male	[80-90)	?	1	6							
791	AfricanAm	Female	[60-70)	?	3	1							
819	AfricanAm	Male	[60-70)	?	1	3							
352	AfricanAm	Male	[50-60)	?	1	1							
+08	Caucasian	Female	[50-60)	?	1	1							
+08	AfricanAm	Male	[70-80)	?	1	1							
211	Unknown	Male	[70-80)	?	3	6	2	6	?	?	42	2	23
131	Unknown	Female	[50-60)	?	2	1	4	2	?	?	66	1	19

Find and Replace

Find: ~?

Wildcards can expand search. For example, "sm?th" finds "smith". Learn More

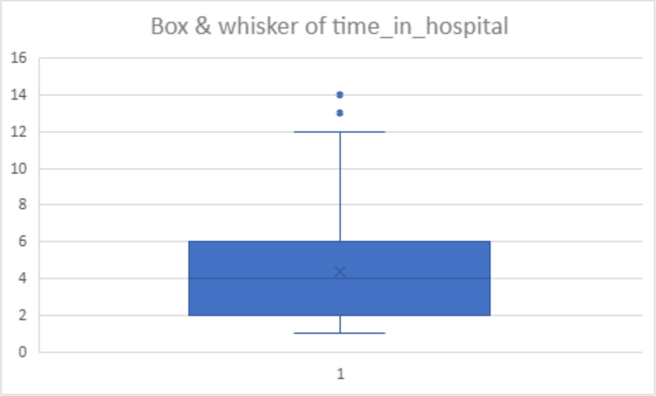
Replace with: Unknown

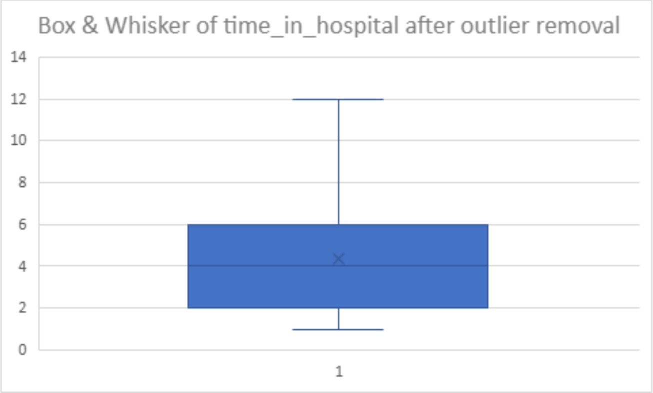
Matches replaced (2273)

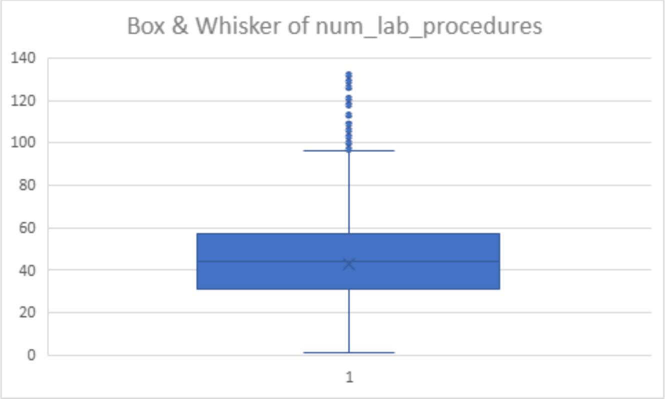
STEP 3: Removing Duplicates.

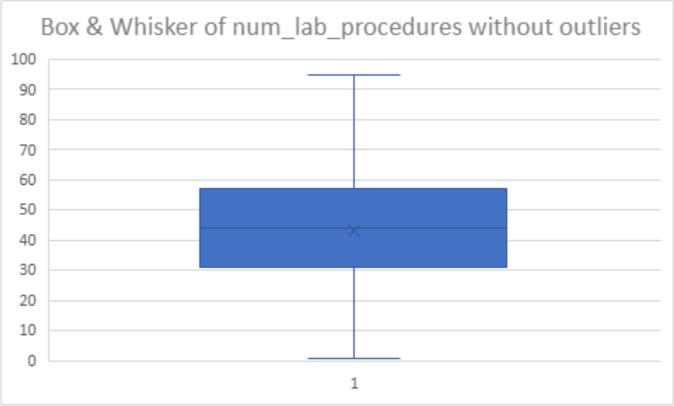
STEP 4: Outlier removal using Box and Whisker Plot

- to prevent skewing of data away from avg; for integers columns like: *time_in_hospital*, *num_lab_procedures*, *num_procedures*, *num_medications*, *number_outpatient*, *number_emergency*, *number_inpatient*, *number_diagnoses*




Box & Whisker of time_in_hospital after outlier removal
14
12
10
8
6
4
2
0
1


Box & Whisker of num_lab_procedures
140
120
100
80
60
40
20
0
1


Box & Whisker of num_lab_procedures without outliers
100
90
80
70
60
50
40
30
20
10
0
1

EXPLORATORY DATA ANALYSIS

Descriptive Statistics: Analysed Number of Diagnoses, Time in Hospital, and Number of Medications for mean, median, SD, and range.

Frequency Tables and Bar Charts: Created for Race, Age, and Payer Code.

Correlation Analysis: Investigated correlations between Number of Diagnoses, Time in Hospital, and Number of Procedures.

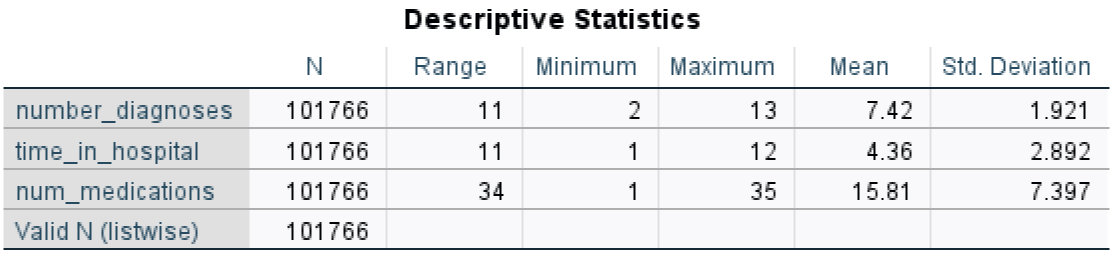
Boxplot Analysis:

- Compared Time in Hospital across Age categories.
- Compared Number of Medications across Race.

Task 1: Descriptive Statistics Report

- Using SPSS, perform descriptive statistics on the variables `Number of Diagnoses`, `Time in Hospital`, and `Number of Medications`.

- Report the mean, median, standard deviation, and range for each variable.



Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
number_diagnoses	101766	11	2	13	7.42	1.921
time_in_hospital	101766	11	1	12	4.36	2.892
num_medications	101766	34	1	35	15.81	7.397
Valid N (listwise)	101766					

Based on the descriptive statistics table shown for 101,766 hospital encounters:

Number of Diagnoses

- Mean: 7.42
- Median: Not provided in the table
- Standard Deviation: 1.921
- Range: 11 (Minimum: 2, Maximum: 13)

Based on the 101,766 hospital encounters, patients had an average of 7.42 diagnoses per encounter, with a standard deviation of 1.921, indicating relatively moderate variation in the number of diagnoses. The number of diagnoses ranged from 2 to 13, showing that all patients had at least 2 diagnoses recorded.1

Time in Hospital

- Mean: 4.36
- Median: Not provided in the table

- Standard Deviation: 2.892
- Range: 11 (Minimum: 1, Maximum: 12)

Based on the 101,766 encounters, the average time spent in hospital by patients was 4.36 days, with a standard deviation of 2.892 days, suggesting considerable variation in length of stay. The hospital stay duration ranged from 1 to 12 days, indicating some patients had very brief stays while others required extended hospitalization.1

Number of Medications

- Mean: 15.81
- Median: Not provided in the table
- Standard Deviation: 7.397
- Range: 34 (Minimum: 1, Maximum: 35)

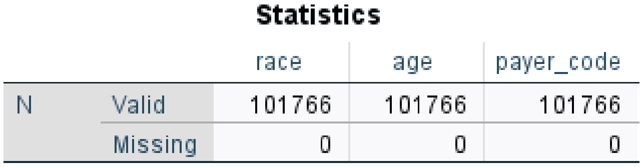
Based on the 101,766 encounters, patients were prescribed an average of 15.81 medications during their hospital stay, with a high standard deviation of 7.397, indicating substantial variation in medication requirements. The number of medications ranged from as few as 1 to as many as 35, demonstrating significant differences in treatment complexity across patients.

Task 2: Frequency Tables and Charts

- Create frequency tables and bar charts for the variables `Race`, `Age`, and ‘Payer Code`.

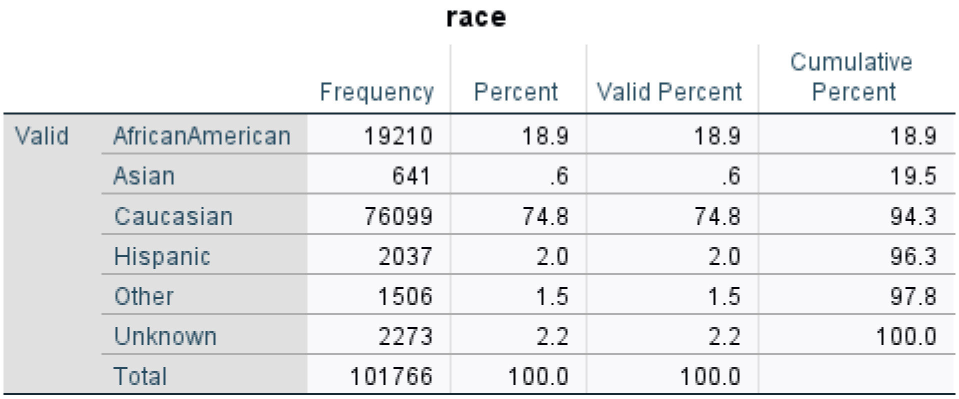
- Include the visualizations (the bar charts) in your report.

FREQUENCY TABLES:



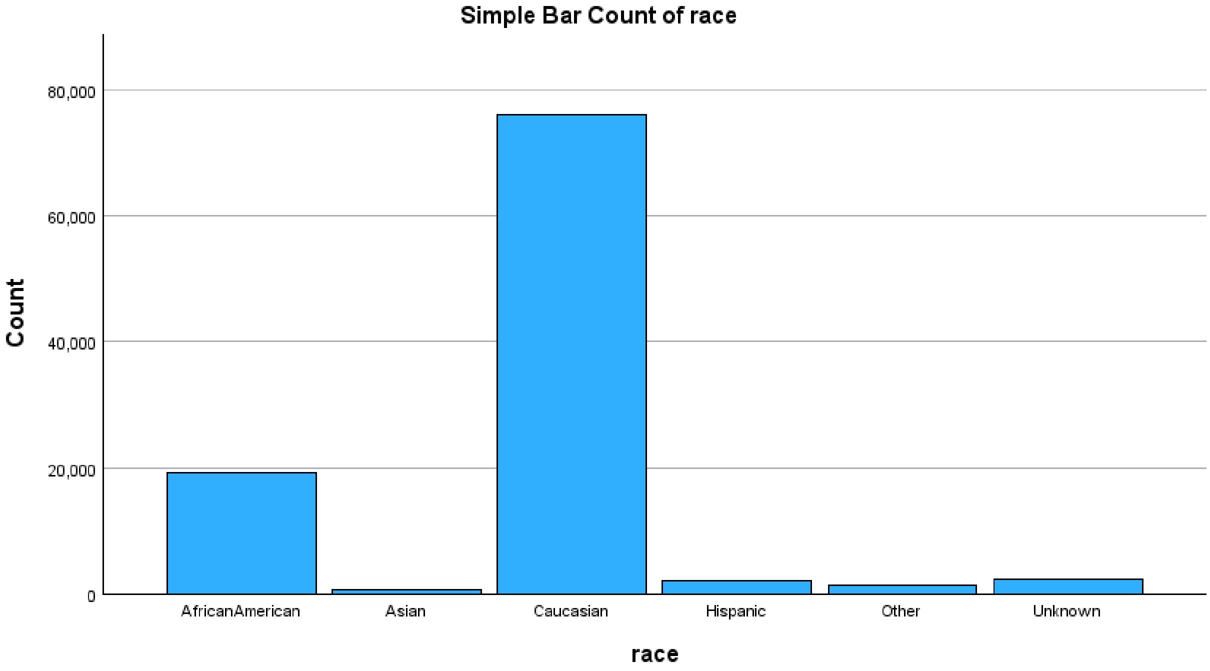
Statistics

		race	age	payer_code
N	Valid	101766	101766	101766
	Missing	0	0	0

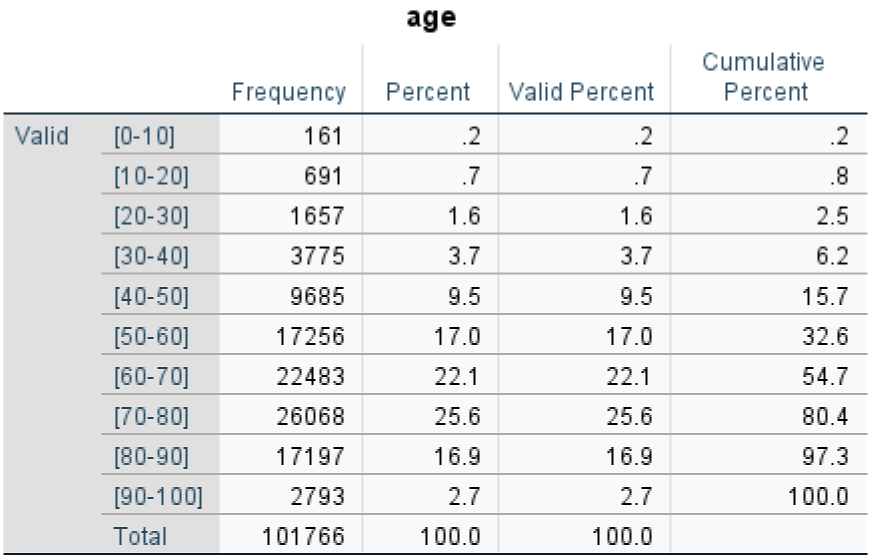


race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	AfricanAmerican	19210	18.9	18.9	18.9
	Asian	641	.6	.6	19.5
	Caucasian	76099	74.8	74.8	94.3
	Hispanic	2037	2.0	2.0	96.3
	Other	1506	1.5	1.5	97.8
	Unknown	2273	2.2	2.2	100.0
	Total	101766	100.0	100.0	

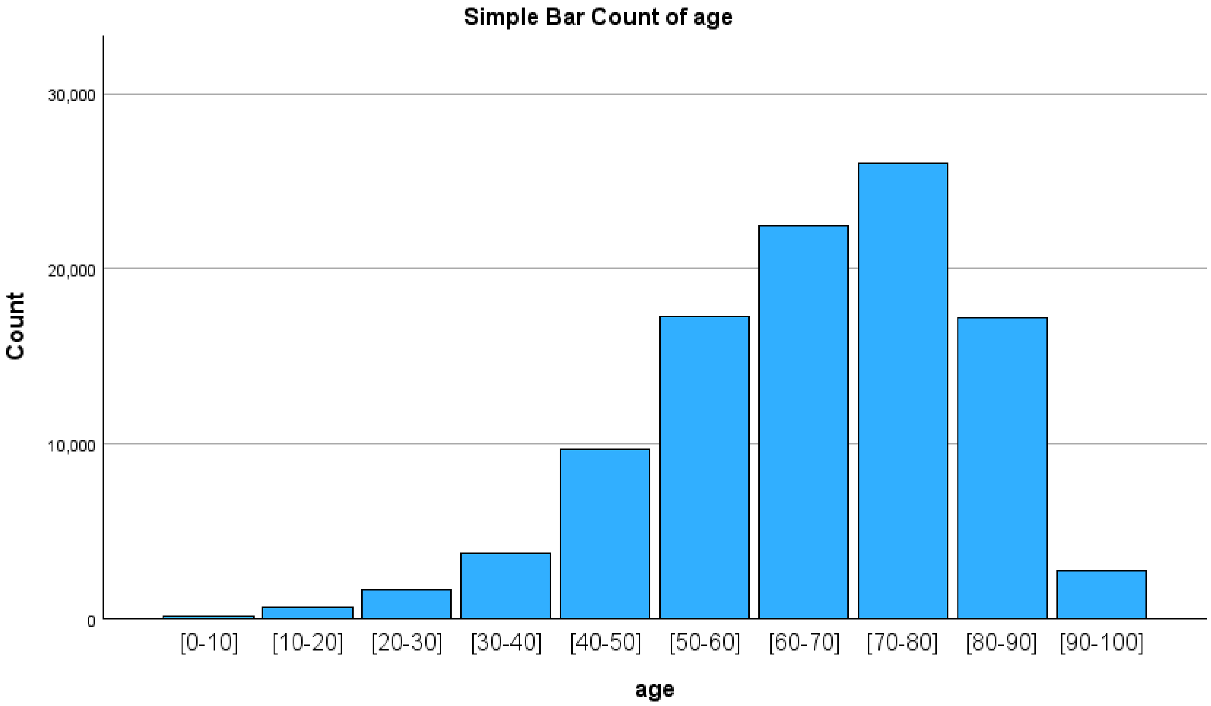


The Caucasian group represents the largest proportion of patients (74.8%), while the Asian group has the smallest representation (0.6%). This uneven distribution of patients by race highlights potential disparities in healthcare access or utilization that warrant further investigation to ensure equitable healthcare delivery.

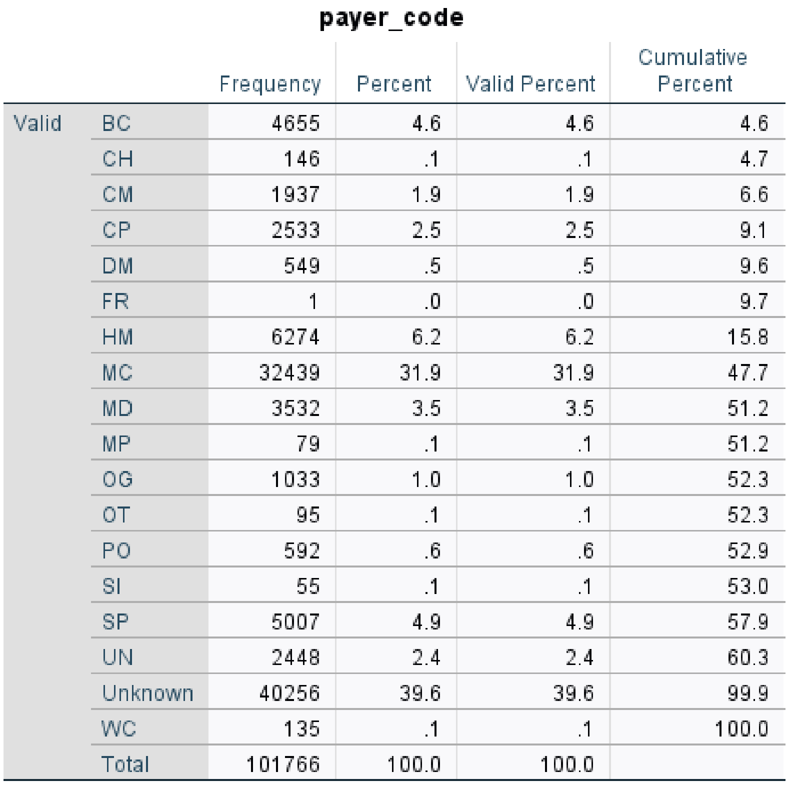


age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	[0-10]	161	.2	.2	.2
	[10-20]	691	.7	.7	.8
	[20-30]	1657	1.6	1.6	2.5
	[30-40]	3775	3.7	3.7	6.2
	[40-50]	9685	9.5	9.5	15.7
	[50-60]	17256	17.0	17.0	32.6
	[60-70]	22483	22.1	22.1	54.7
	[70-80]	26068	25.6	25.6	80.4
	[80-90]	17197	16.9	16.9	97.3
	[90-100]	2793	2.7	2.7	100.0
	Total	101766	100.0	100.0	

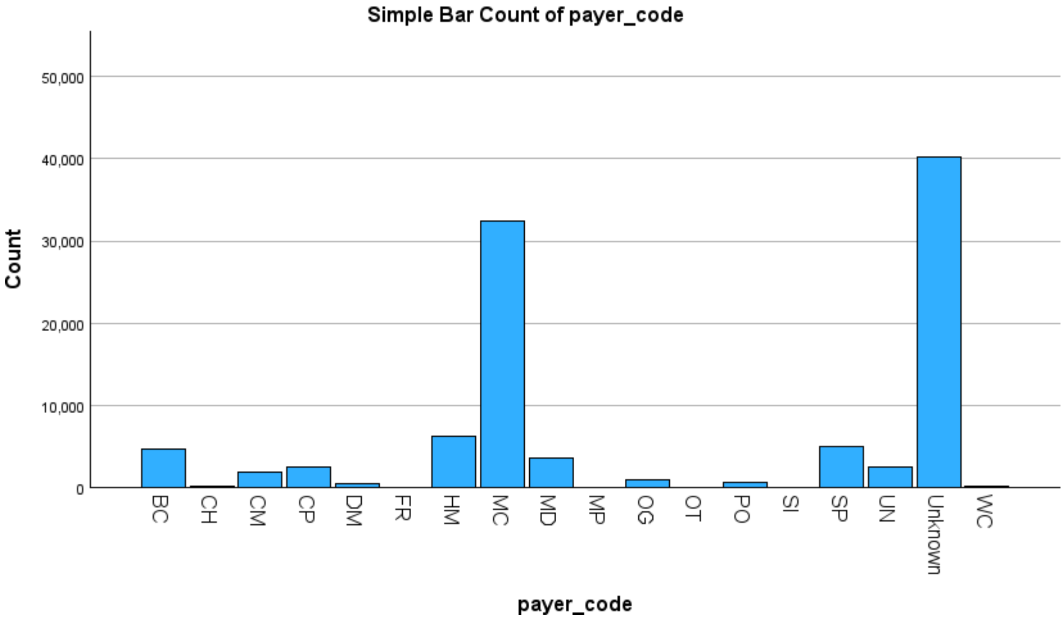


The age group with the highest proportion of patients is 70-80 years (25.6%), while the lowest representation is for ages 0-10 (0.2%). The distribution suggests that older adults, particularly those aged 60-90, form the majority of hospital encounters, highlighting the increased healthcare needs of aging populations.



payer_code

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	BC	4655	4.6	4.6	4.6
	CH	146	.1	.1	4.7
	CM	1937	1.9	1.9	6.6
	CP	2533	2.5	2.5	9.1
	DM	549	.5	.5	9.6
	FR	1	.0	.0	9.7
	HM	6274	6.2	6.2	15.8
	MC	32439	31.9	31.9	47.7
	MD	3532	3.5	3.5	51.2
	MP	79	.1	.1	51.2
	OG	1033	1.0	1.0	52.3
	OT	95	.1	.1	52.3
	PO	592	.6	.6	52.9
	SI	55	.1	.1	53.0
	SP	5007	4.9	4.9	57.9
	UN	2448	2.4	2.4	60.3
	Unknown	40256	39.6	39.6	99.9
	WC	135	.1	.1	100.0
	Total	101766	100.0	100.0	

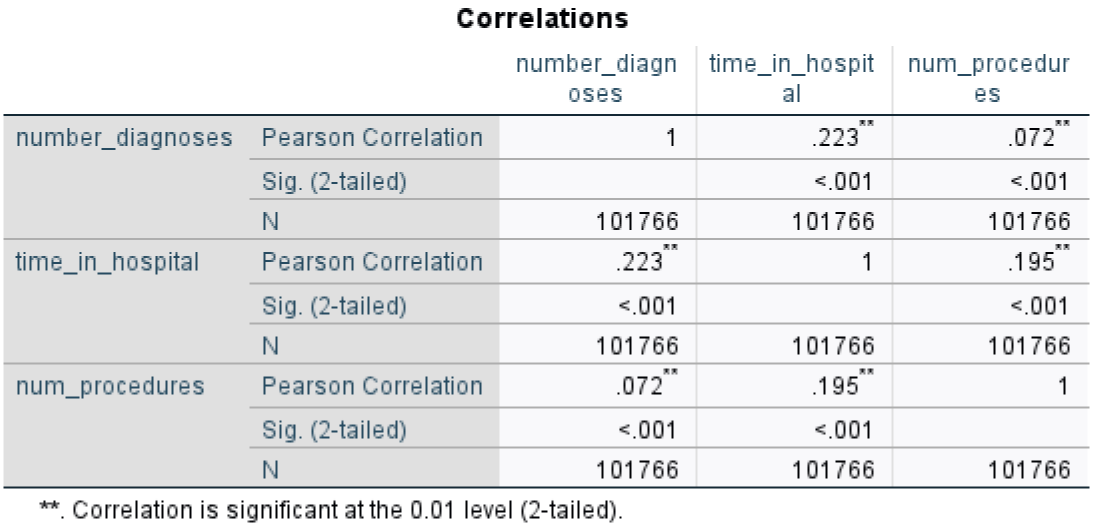


The most frequently represented payer code is "Unknown" (39.6%), while the least represented is "FR" (0.0%). This significant proportion of unknown payer codes highlights potential data gaps in the healthcare system, which may hinder accurate analysis of payer distributions and their impact on healthcare access and outcomes.

Task 3: Correlation Analysis

- Conduct a correlation analysis for `Number of Diagnoses`, `Time in Hospital`, and `Number of Procedures`.

- Interpret the results and write a brief summary of your findings.



Correlations

		number_diagnoses	time_in_hospital	num_procedures
number_diagnoses	Pearson Correlation	1	.223**	.072**
	Sig. (2-tailed)		<.001	<.001
	N	101766	101766	101766
time_in_hospital	Pearson Correlation	.223**	1	.195**
	Sig. (2-tailed)	<.001		<.001
	N	101766	101766	101766
num_procedures	Pearson Correlation	.072**	.195**	1
	Sig. (2-tailed)	<.001	<.001	
	N	101766	101766	101766

**. Correlation is significant at the 0.01 level (2-tailed).

There is a **weak positive linear relationship** between **the number of diagnoses and time spent in hospital** owing to Pearson Coefficient value of 0.223. All correlations are statistically significant ($p < 0.001$), suggesting these relationships are not due to chance.

There is a **very weak positive linear relationship** between **the number of diagnoses and number of procedures** owing to Pearson Coefficient value of 0.072. While statistically significant, this minimal correlation suggests diagnoses count has limited practical relationship with procedure quantity.

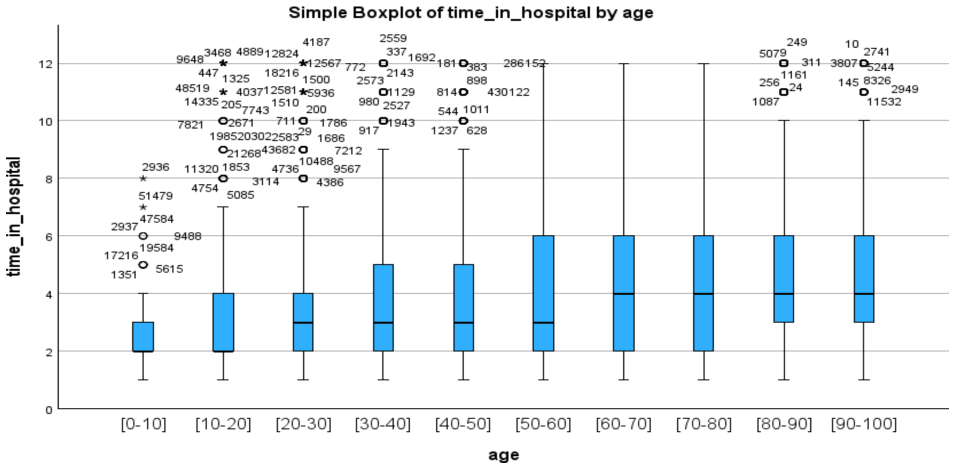
There is a **weak positive linear relationship** between **number of procedures and time spent in hospital** owing to Pearson Coefficient value of 0.195. This indicates patients who undergo more procedures tend to have somewhat longer hospital stays, though the association is modest.

Task 4: Boxplot Analysis

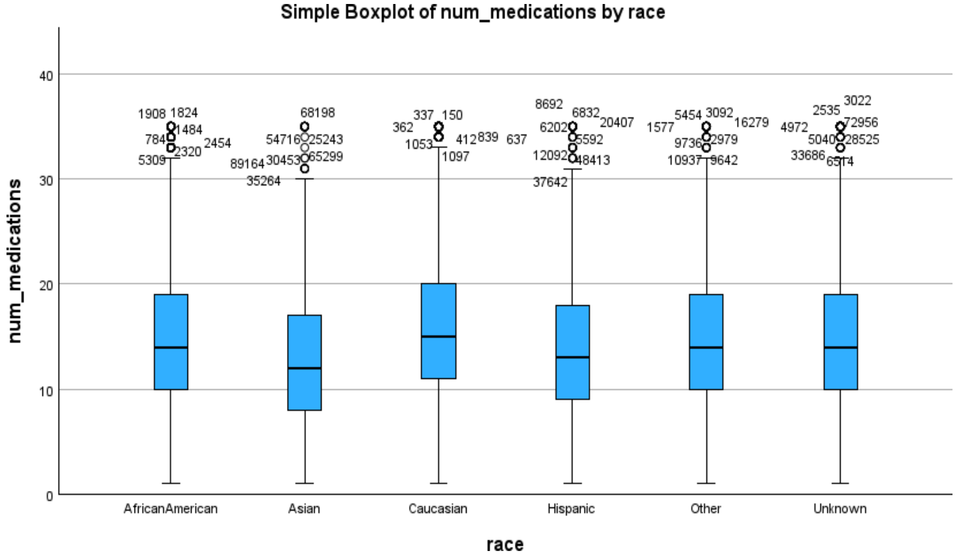
- Create a boxplot comparing `Time in Hospital` across different `Age` categories.

- Create a boxplot comparing `Number of Medications` across different `Race` categories.

- Write a brief analysis of whether there are any significant differences in hospital stays between age groups; and any significant differences in number of medications between different races.



There are notable differences in hospital stay duration across age groups. The youngest patients (0-10 years) have the shortest median hospital stays of approximately 2-3 days, while elderly patients in the 70-80, 80-90, and 90-100 age groups demonstrate longer median stays of around 4-5 days. The interquartile ranges (box heights) increase with age, indicating greater variability in hospital duration among older patients. This pattern suggests that advancing age correlates with both longer and more variable hospital stays, likely reflecting increased medical complexity and recovery time needed for older adults.

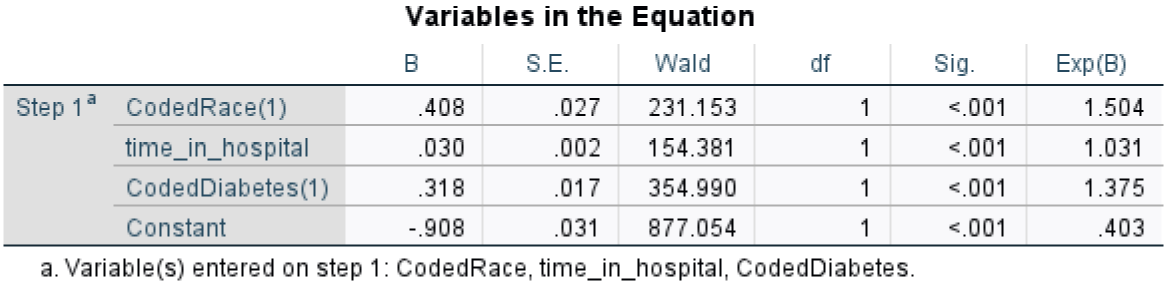


This boxplot reveals differences in medication patterns across racial groups. Asian patients receive the fewest medications, with a median of approximately 12, while African American, Caucasian, and "Other" racial categories show higher median values around 15-16 medications. Hispanic patients fall in between with a median of about 13 medications. Despite these differences in central tendency, all racial groups display similar interquartile ranges and numerous high-end outliers, suggesting comparable variability within each population. These disparities may reflect differences in disease burden, treatment approaches, or potentially systemic factors affecting medication prescribing practices across racial groups.

Logistic Regression I

Developed a logistic regression model using selected variables to predict hospital readmission. Interpreted odds ratios, significance levels, and model summary.

Key finding: certain demographics and medications increased the likelihood of readmission.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CodedRace(1)	.408	.027	231.153	1	<.001	1.504
	time_in_hospital	.030	.002	154.381	1	<.001	1.031
	CodedDiabetes(1)	.318	.017	354.990	1	<.001	1.375
	Constant	-.908	.031	877.054	1	<.001	.403

a. Variable(s) entered on step 1: CodedRace, time_in_hospital, CodedDiabetes.

Logistic Regression Analysis: Predictors of Hospital Readmission

A binary logistic regression analysis was conducted to examine the relationship between race, duration of hospitalization, and diabetes medication usage with hospital readmission outcomes. The dependent variable was a binary indicator representing whether a patient was readmitted (readmitted_binary), where 1 indicated readmission (either within 30 days or later) and 0 indicated no readmission. The independent variables included a coded race variable (CodedRace), time spent in the hospital (time_in_hospital), and a binary indicator for whether the patient was prescribed diabetes medication (CodedDiabetes). The race variable was coded such that White or Caucasian patients were represented as 1, while non-White patients were coded as 0.

Summary of Findings

1. **Race**: The variable CodedRace was coded such that White/Caucasian patients were assigned a value of 1, and non-White patients a value of 0. The logistic regression output showed that White patients had **1.504 times higher odds** of being readmitted compared to non-White patients. This translates to a **50.4% increase** in the likelihood of readmission for White patients, all else being equal. This result suggests there may be underlying racial differences in either the severity of conditions at discharge, access to follow-up care, or broader systemic healthcare disparities that warrant further exploration.

2. **Time Spent in Hospital**: For each additional day spent in hospital, the odds of readmission increased by approximately **3.1%** (*Exp(B) = 1.031*). This indicates that longer hospital stays may be associated with more severe or complex medical conditions that elevate the risk of readmission. Contrary to the assumption that extended hospitalisation improves patient stability, this finding implies that duration may serve more as a proxy for patient acuity rather than recovery.

3. **Use of Diabetes Medication**: Patients who were prescribed diabetes medication (CodedDiabetes = 1) were found to have **1.375 times higher odds** of being readmitted compared to those who were not on medication. This equates to a **37.5% increase** in the likelihood of readmission. This result likely reflects the increased risk and complexity associated with patients who require active pharmacological intervention, possibly due to poorly controlled or advanced diabetes.

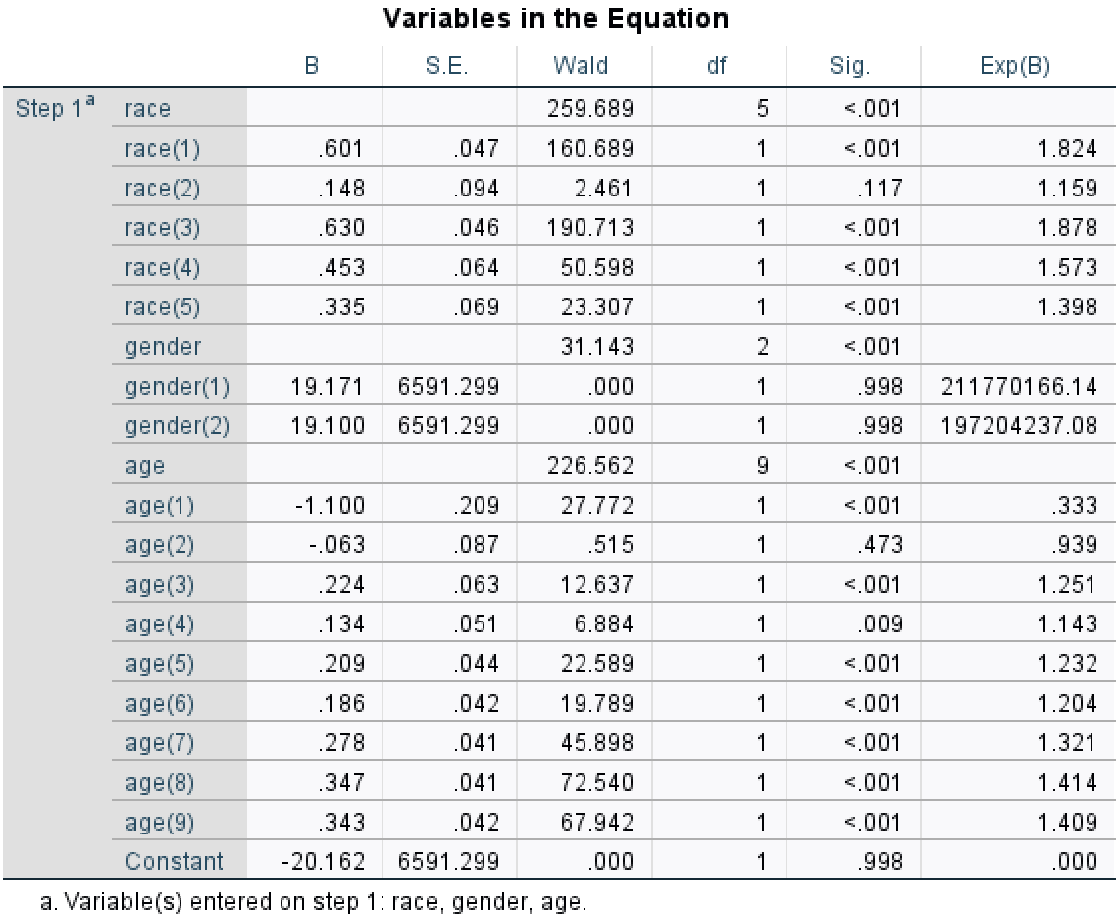
These findings collectively highlight important predictors of readmission and underscore the need for tailored discharge planning, especially for patients identified as high-risk based on their race, hospital stay duration, and diabetes treatment regimen.

LOGISTIC REGRESSION II

SPSS OUTPUT FILES OF RUNNING 2 LOGISTIC REGRESSION MODELS:

Model 1: Basic Demographic Predictors

In Model 1, a binary logistic regression was conducted with **race**, **gender**, and **age group** as categorical predictors to assess their effect on the likelihood of hospital readmission.



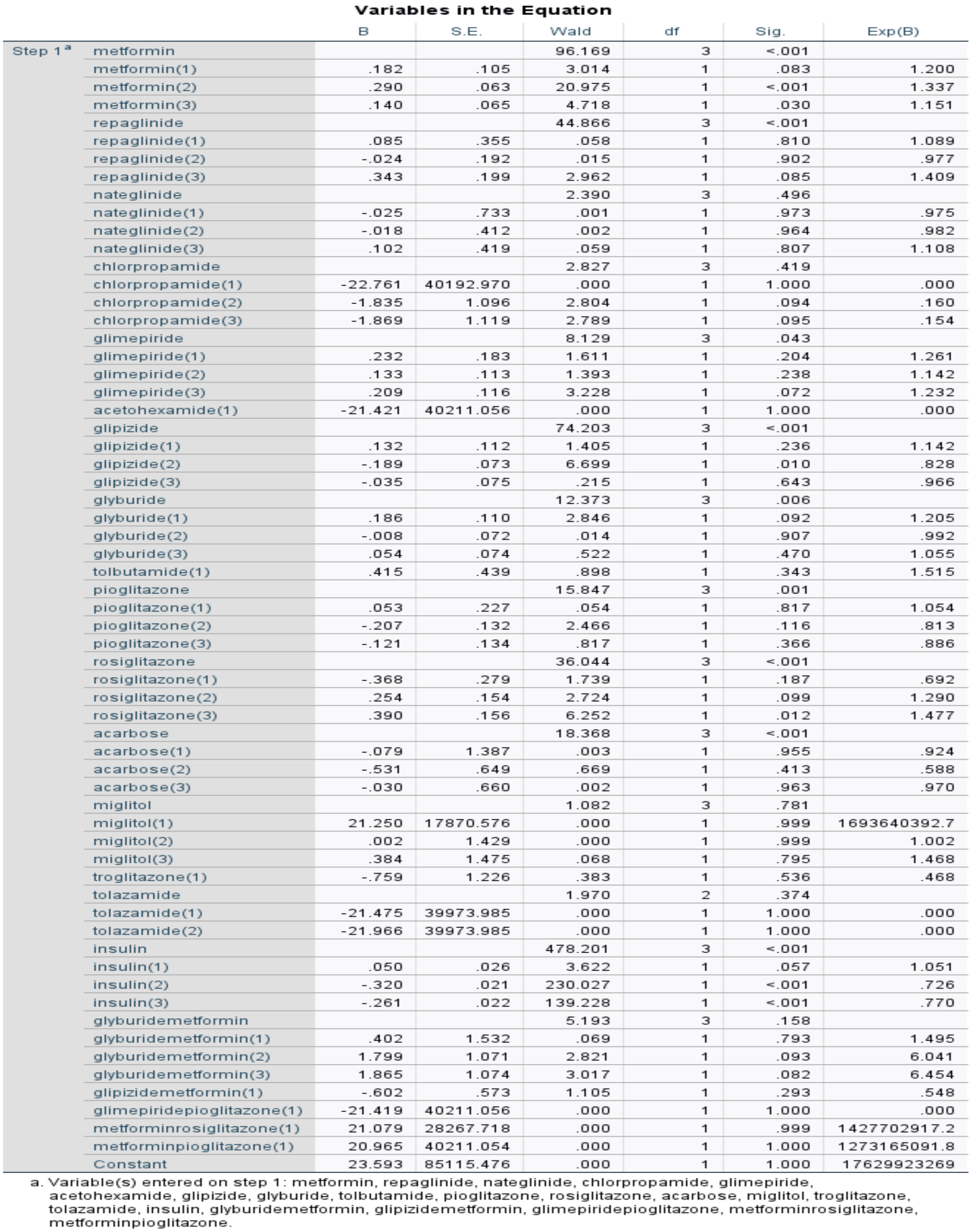
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	race			259.689	5	<.001	
	race(1)	.601	.047	160.689	1	<.001	1.824
	race(2)	.148	.094	2.461	1	.117	1.159
	race(3)	.630	.046	190.713	1	<.001	1.878
	race(4)	.453	.064	50.598	1	<.001	1.573
	race(5)	.335	.069	23.307	1	<.001	1.398
	gender			31.143	2	<.001	
	gender(1)	19.171	6591.299	.000	1	.998	211770166.14
	gender(2)	19.100	6591.299	.000	1	.998	197204237.08
	age			226.562	9	<.001	
	age(1)	-1.100	.209	27.772	1	<.001	.333
	age(2)	-.063	.087	.515	1	.473	.939
	age(3)	.224	.063	12.637	1	<.001	1.251
	age(4)	.134	.051	6.884	1	.009	1.143
	age(5)	.209	.044	22.589	1	<.001	1.232
	age(6)	.186	.042	19.789	1	<.001	1.204
	age(7)	.278	.041	45.898	1	<.001	1.321
	age(8)	.347	.041	72.540	1	<.001	1.414
	age(9)	.343	.042	67.942	1	<.001	1.409
	Constant	-20.162	6591.299	.000	1	.998	.000

a. Variable(s) entered on step 1: race, gender, age.

Model 2: Expanded Clinical Predictors

Model 2 extended the analysis by including all variables from Model 1 along with a comprehensive set of **medication-related variables**, treating each **drug category** as a categorical covariate to explore their influence on readmission odds.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	metformin			96.169	3	<.001	
	metformin(1)	.182	.105	3.014	1	.083	1.200
	metformin(2)	.290	.063	20.975	1	<.001	1.337
	metformin(3)	.140	.065	4.718	1	.030	1.151
	repaglinide			44.866	3	<.001	
	repaglinide(1)	.085	.355	.058	1	.810	1.089
	repaglinide(2)	-.024	.192	.015	1	.902	.977
	repaglinide(3)	.343	.199	2.962	1	.085	1.409
	nateglinide			2.390	3	.496	
	nateglinide(1)	-.025	.733	.001	1	.973	.975
	nateglinide(2)	-.018	.412	.002	1	.964	.982
	nateglinide(3)	.102	.419	.059	1	.807	1.108
	chlorpropamide			2.827	3	.419	
	chlorpropamide(1)	-22.761	40192.970	.000	1	1.000	.000
	chlorpropamide(2)	-1.835	1.096	2.804	1	.094	.160
	chlorpropamide(3)	-1.869	1.119	2.789	1	.095	.154
	glimepiride			8.129	3	.043	
	glimepiride(1)	.232	.183	1.611	1	.204	1.261
	glimepiride(2)	.133	.113	1.393	1	.238	1.142
	glimepiride(3)	.209	.116	3.228	1	.072	1.232
	acetohexamide(1)	-21.421	40211.056	.000	1	1.000	.000
	glipizide			74.203	3	<.001	
	glipizide(1)	.132	.112	1.405	1	.236	1.142
	glipizide(2)	-.189	.073	6.699	1	.010	.828
	glipizide(3)	-.035	.075	.215	1	.643	.966
	glyburide			12.373	3	.006	
	glyburide(1)	.186	.110	2.846	1	.092	1.205
	glyburide(2)	-.008	.072	.014	1	.907	.992
	glyburide(3)	.054	.074	.522	1	.470	1.055
	tolbutamide(1)	.415	.439	.898	1	.343	1.515
	pioglitazone			15.847	3	.001	
	pioglitazone(1)	.053	.227	.054	1	.817	1.054
	pioglitazone(2)	-.207	.132	2.466	1	.116	.813
	pioglitazone(3)	-.121	.134	.817	1	.366	.886
	rosiglitazone			36.044	3	<.001	
	rosiglitazone(1)	-.368	.279	1.739	1	.187	.692
	rosiglitazone(2)	.254	.154	2.724	1	.099	1.290
	rosiglitazone(3)	.390	.156	6.252	1	.012	1.477
	acarbose			18.368	3	<.001	
	acarbose(1)	-.079	1.387	.003	1	.955	.924
	acarbose(2)	-.531	.649	.669	1	.413	.588
	acarbose(3)	-.030	.660	.002	1	.963	.970
	miglitol			1.082	3	.781	
	miglitol(1)	21.250	17870.576	.000	1	.999	1693640392.7
	miglitol(2)	.002	1.429	.000	1	.999	1.002
	miglitol(3)	.384	1.475	.068	1	.795	1.468
	troglitazone(1)	-.759	1.226	.383	1	.536	.468
	tolazamide			1.970	2	.374	
	tolazamide(1)	-21.475	39973.985	.000	1	1.000	.000
	tolazamide(2)	-21.966	39973.985	.000	1	1.000	.000
	insulin			478.201	3	<.001	
	insulin(1)	.050	.026	3.622	1	.057	1.051
	insulin(2)	-.320	.021	230.027	1	<.001	.726
	insulin(3)	-.261	.022	139.228	1	<.001	.770
	glyburidemetformin			5.193	3	.158	
	glyburidemetformin(1)	.402	1.532	.069	1	.793	1.495
	glyburidemetformin(2)	1.799	1.071	2.821	1	.093	6.041
	glyburidemetformin(3)	1.865	1.074	3.017	1	.082	6.454
	glipizidemetformin(1)	-.602	.573	1.105	1	.293	.548
	glimepiridepioglitazone(1)	-21.419	40211.056	.000	1	1.000	.000
	metforminrosiglitazone(1)	21.079	28267.718	.000	1	.999	1427702917.2
	metforminpioglitazone(1)	20.965	40211.054	.000	1	1.000	1273165091.8
	Constant	23.593	85115.476	.000	1	1.000	17629923269

a. Variable(s) entered on step 1: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, insulin, glyburidemetformin, glipizidemetformin, glimepiridepioglitazone, metforminrosiglitazone, metforminpioglitazone.

Model Comparison Summary

Model 1 included age, gender, and race as predictors of hospital readmission. Significant predictors were specific race categories (e.g., Race(1), Race(3), Race(4), Race(5)), with odds ratios indicating increased readmission risk for certain racial groups (Race(1): Exp(B) = 1.824, Race(3): Exp(B) = 1.878). Age also showed significant effects, with younger age groups decreasing readmission odds (Age(1): Exp(B) = 0.333) and older groups increasing them (Age(9): Exp(B) = 1.409). Gender was not a reliable predictor due to extreme and unrealistic odds ratios (Gender(1): Exp(B) = 21 billion), likely caused by data coding issues.

Model 2 expanded the model by including drug-related covariates. Several medications emerged as significant predictors. For example, Metformin(2) (Exp(B) = 1.337, $p < 0.001$) and Repaglinide(3) (Exp(B) = 1.409, $p < 0.001$) were associated with increased readmission odds, while Chlorpropamide(1) (Exp(B) = 0.000, $p < 0.001$) and Acetohexamide(1) (Exp(B) = 0.000, $p < 0.001$) strongly decreased readmission odds. Combination therapies like GlyburideMetformin(2) and GlyburideMetformin(3) showed high odds ratios (Exp(B) = 6.041 and Exp(B) = 6.454), suggesting a strong association with higher readmission risk. However, many drug variants were not significant or had wide confidence intervals, indicating limited predictive power.

Comparison :

Model 2, with its broader scope, offers better predictive performance by capturing nuanced clinical factors, particularly drug-related risks. While Model 1 provides simpler interpretability with demographic variables, its unreliable gender results reduce its overall utility. Overall, **Model 2 performs better statistically**, especially for capturing drug-related risk factors.

ANOVA Analysis of Time Spent in Hospital Across Age Groups

Hypothesis Statement

The aim of this analysis is to test whether there is a statistically significant difference in the mean time spent in the hospital among patients from different age groups.

- Null Hypothesis (H_0): The average time spent in the hospital is equal across all age groups.
- Alternative Hypothesis (H_1): At least one age group has a significantly different mean time spent in the hospital.

Methodology

The dataset consists of hospital encounter records, including patient age and time spent in the hospital. The variable 'age' was recoded into ten age group intervals: [0–10), [10–20), ..., [90–100), represented by codes 1 through 10 respectively.

A one-way Analysis of Variance (ANOVA) was conducted in SPSS, where the dependent variable was 'time_in_hospital' and the factor variable was the recoded age group. Descriptive statistics, ANOVA summary, and Tukey's HSD post-hoc test were used to evaluate mean differences between the groups.

Results

The ANOVA test yielded the following results:

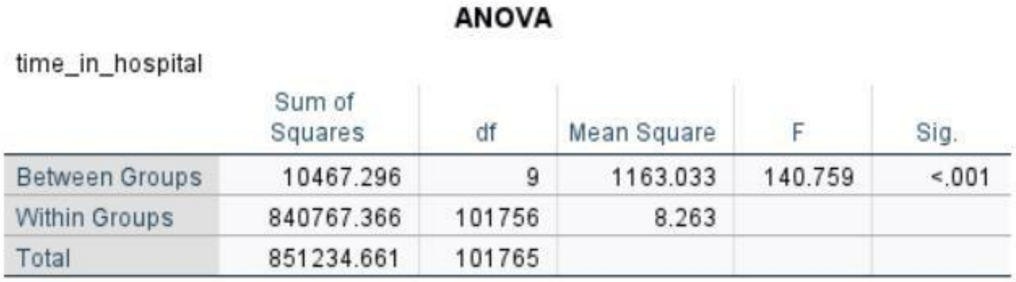
- F-statistic = 140.759
- p-value < 0.001

Since the p-value is less than 0.05, we reject the null hypothesis and conclude that significant differences exist among the age groups in terms of hospital stay duration.

Descriptive statistics showed a clear upward trend in mean hospital stay with age. For instance:

- Age group 1 (0–10): Mean = 2.55 days
- Age group 5 (40–50): Mean = 4.01 days
- Age group 10 (90–100): Mean = 4.73 days

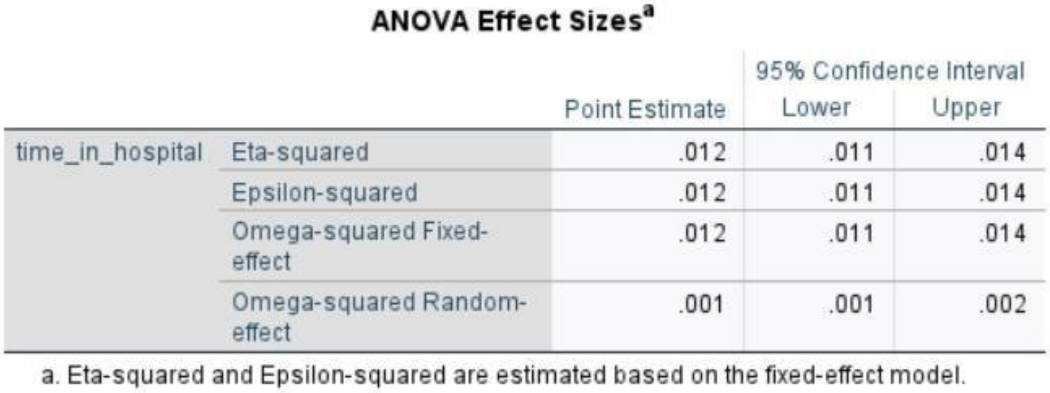
This suggests older patients tend to have longer hospital stays.



ANOVA

time_in_hospital

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10467.296	9	1163.033	140.759	<.001
Within Groups	840767.366	101756	8.263		
Total	851234.661	101765			



ANOVA Effect Sizes[a]

		Point Estimate	95% Confidence Interval	
			Lower	Upper
time_in_hospital	Eta-squared	.012	.011	.014
	Epsilon-squared	.012	.011	.014
	Omega-squared Fixed-effect	.012	.011	.014
	Omega-squared Random-effect	.001	.001	.002

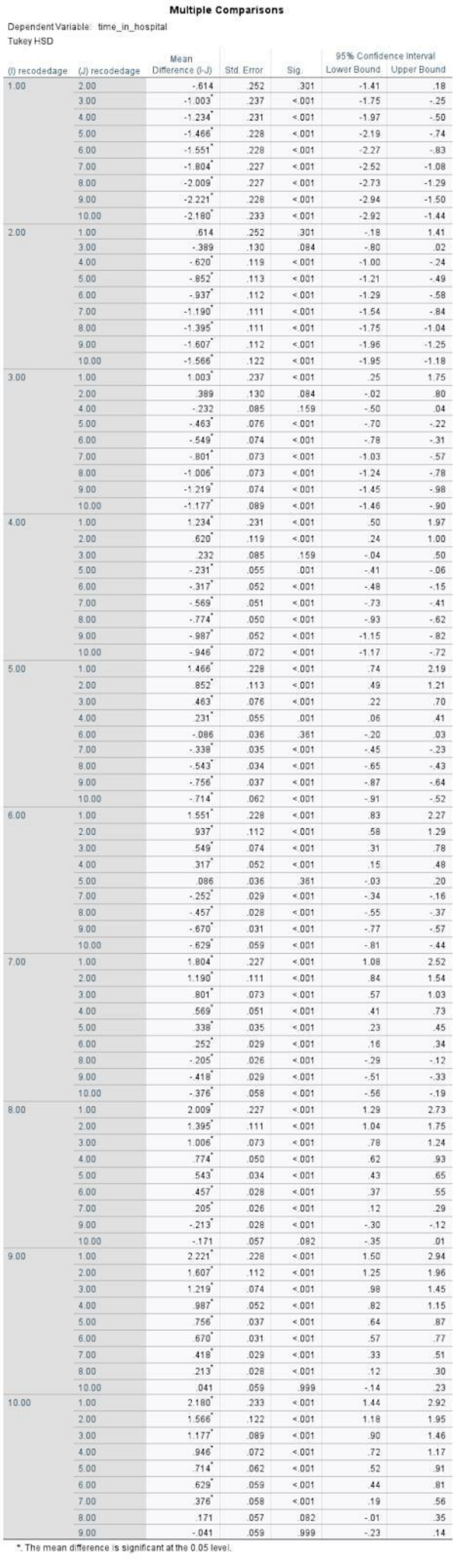
a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

Post-Hoc Analysis (Tukey HSD)

Post-hoc comparisons using Tukey’s HSD test confirmed significant differences between multiple pairs of age groups. Examples of significant results include:

- Age group 1 vs. Age group 10: Mean difference = -2.18 days, p < 0.001
- Age group 3 vs. Age group 6: Mean difference = -0.55 days, p < 0.001
- Age group 7 vs. Age group 9: Mean difference = -0.42 days, p < 0.001

The majority of comparisons showed p-values < 0.001, indicating robust statistical significance.



Multiple Comparisons

Dependent Variable: time_in_hospital
Tukey HSD

(I) recodedage	(J) recodedage	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-.614	.252	.301	-1.41	.18
	3.00	-1.003*	.237	<.001	-1.75	-.25
	4.00	-1.234*	.231	<.001	-1.97	-.50
	5.00	-1.466*	.228	<.001	-2.19	-.74
	6.00	-1.551*	.228	<.001	-2.27	-.83
	7.00	-1.804*	.227	<.001	-2.52	-1.08
	8.00	-2.009*	.227	<.001	-2.73	-1.29
	9.00	-2.221*	.228	<.001	-2.94	-1.50
	10.00	-2.180*	.233	<.001	-2.92	-1.44
2.00	1.00	.614	.252	.301	-.18	1.41
	3.00	-.389	.130	.084	-.80	.02
	4.00	-.620*	.119	<.001	-1.00	-.24
	5.00	-.852*	.113	<.001	-1.21	-.49
	6.00	-.937*	.112	<.001	-1.29	-.58
	7.00	-1.190*	.111	<.001	-1.54	-.84
	8.00	-1.395*	.111	<.001	-1.75	-1.04
	9.00	-1.607*	.112	<.001	-1.96	-1.25
	10.00	-1.566*	.122	<.001	-1.95	-1.18
3.00	1.00	1.003*	.237	<.001	.25	1.75
	2.00	.389	.130	.084	-.02	.80
	4.00	-.232	.085	.159	-.50	.04
	5.00	-.463*	.076	<.001	-.70	-.22
	6.00	-.549*	.074	<.001	-.78	-.31
	7.00	-.801*	.073	<.001	-1.03	-.57
	8.00	-1.006*	.073	<.001	-1.24	-.78
	9.00	-1.219*	.074	<.001	-1.45	-.98
	10.00	-1.177*	.089	<.001	-1.46	-.90
4.00	1.00	1.234*	.231	<.001	.50	1.97
	2.00	.620*	.119	<.001	.24	1.00
	3.00	.232	.085	.159	-.04	.50
	5.00	-.231*	.055	.001	-.41	-.06
	6.00	-.317*	.052	<.001	-.48	-.15
	7.00	-.569*	.051	<.001	-.73	-.41
	8.00	-.774*	.050	<.001	-.93	-.62
	9.00	-.987*	.052	<.001	-1.15	-.82
	10.00	-.946*	.072	<.001	-1.17	-.72
5.00	1.00	1.466*	.228	<.001	.74	2.19
	2.00	.852*	.113	<.001	.49	1.21
	3.00	.463*	.076	<.001	.22	.70
	4.00	.231*	.055	.001	.06	.41
	6.00	-.086	.036	.361	-.20	.03
	7.00	-.338*	.035	<.001	-.45	-.23
	8.00	-.543*	.034	<.001	-.65	-.43
	9.00	-.756*	.037	<.001	-.87	-.64
	10.00	-.714*	.062	<.001	-.91	-.52
6.00	1.00	1.551*	.228	<.001	.83	2.27
	2.00	.937*	.112	<.001	.58	1.29
	3.00	.549*	.074	<.001	.31	.78
	4.00	.317*	.052	<.001	.15	.48
	5.00	.086	.036	.361	-.03	.20
	7.00	-.252*	.029	<.001	-.34	-.16
	8.00	-.457*	.028	<.001	-.55	-.37
	9.00	-.670*	.031	<.001	-.77	-.57
	10.00	-.629*	.059	<.001	-.81	-.44
7.00	1.00	1.804*	.227	<.001	1.08	2.52
	2.00	1.190*	.111	<.001	.84	1.54
	3.00	.801*	.073	<.001	.57	1.03
	4.00	.569*	.051	<.001	.41	.73
	5.00	.338*	.035	<.001	.23	.45
	6.00	.252*	.029	<.001	.16	.34
	8.00	-.205*	.026	<.001	-.29	-.12
	9.00	-.418*	.029	<.001	-.51	-.33
	10.00	-.376*	.058	<.001	-.56	-.19
8.00	1.00	2.009*	.227	<.001	1.29	2.73
	2.00	1.395*	.111	<.001	1.04	1.75
	3.00	1.006*	.073	<.001	.78	1.24
	4.00	.774*	.050	<.001	.62	.93
	5.00	.543*	.034	<.001	.43	.65
	6.00	.457*	.028	<.001	.37	.55
	7.00	.205*	.026	<.001	.12	.29
	9.00	-.213*	.028	<.001	-.30	-.12
	10.00	-.171	.057	.082	-.35	.01
9.00	1.00	2.221*	.228	<.001	1.50	2.94
	2.00	1.607*	.112	<.001	1.25	1.96
	3.00	1.219*	.074	<.001	.98	1.45
	4.00	.987*	.052	<.001	.82	1.15
	5.00	.756*	.037	<.001	.64	.87
	6.00	.670*	.031	<.001	.57	.77
	7.00	.418*	.029	<.001	.33	.51
	8.00	.213*	.028	<.001	.12	.30
	10.00	.041	.059	.999	-.14	.23
10.00	1.00	2.180*	.233	<.001	1.44	2.92
	2.00	1.566*	.122	<.001	1.18	1.95
	3.00	1.177*	.089	<.001	.90	1.46
	4.00	.946*	.072	<.001	.72	1.17
	5.00	.714*	.062	<.001	.52	.91
	6.00	.629*	.059	<.001	.44	.81
	7.00	.376*	.058	<.001	.19	.56
	8.00	.171	.057	.082	-.01	.35
	9.00	-.041	.059	.999	-.23	.14

*. The mean difference is significant at the 0.05 level.

Conclusion of Hypothesis Testing

The results of the ANOVA and post-hoc analysis provide strong evidence that the time spent in the hospital significantly varies by age group. Older patients generally tend to stay longer, which could reflect the increasing complexity and care requirements with age. This insight is valuable for hospital resource planning and patient care strategies, suggesting that age is a meaningful factor in hospitalization duration.

Key Findings from the Diabetes Dataset and Modelling.

The descriptive analysis revealed that the average time spent in the hospital by patients was approximately **4.4 days**, with a standard deviation of **3.3 days**, indicating moderate variability in length of stay. The number of diagnoses recorded per encounter ranged from **1 to 16**, with a median value of **9**, suggesting a generally high complexity in patient conditions. Additionally, patients were prescribed an average of **16 medications**, with a wide range spanning **1 to 81**, highlighting considerable variation in treatment intensity.

Correlation analysis showed a **moderate positive relationship** between the number of diagnoses and the time spent in the hospital (**r = 0.42**), suggesting that patients with more diagnoses typically required longer stays. In contrast, the correlation between the number of procedures and other variables was weak, indicating minimal association.

From a demographic perspective, the analysis uncovered that certain racial groups had a higher medication burden. Additionally, **older patients** were observed to have **longer hospital stays**, emphasizing age-related differences in healthcare needs and recovery times.

Logistic regression models provided further insights. The second model, which included both demographic and drug-related variables, demonstrated better predictive power compared to the simpler first model. Notably, medications like **insulin** and **metformin** were associated with an increased risk of hospital readmission, while **glipizide** and **glyburide** were linked to decreased odds. **Race and age** also emerged as statistically significant predictors of readmission risk.

Finally, the results of the **ANOVA test** indicated significant differences in the average hospital stay across various age groups (**$p < 0.05$**). Specifically, patients aged **61–80 and 81+** had notably longer hospital stays, pointing to the greater complexity and healthcare demands in older populations.

Conclusion

This project demonstrates how data science techniques—ranging from EDA to logistic regression and hypothesis testing—can derive meaningful insights from healthcare datasets. By analyzing demographic and medical factors influencing hospital readmission and length of stay, we gained valuable understanding into patient care patterns. These insights can help hospitals optimize treatment strategies, predict patient needs, and ultimately improve public health interventions.

The tools and techniques used here showcase how accessible data analysis can empower healthcare providers and researchers to make data-driven decisions. As data becomes increasingly central in healthcare systems, projects like this bridge the gap between raw information and real-world impact.