dataset

September 5, 2024

HEALTHCARE DATASET

Loading the Dataset:

[ ]:
```
import pandas as pd
import numpy as np
from scipy import stats
import statsmodels.api as sm

# Load the dataset
df_new = pd.read_csv('/content/healthcare_dataset.csv')

# Display the first few rows to inspect the data
print(df_new.head())
```

```
            Name  Age  Gender Blood Type Medical Condition Date of Admission  \
0  Bobby JacksOn   30    Male         B-            Cancer        2024-01-31
1   LesLie TErRy   62    Male         A+           Obesity        2019-08-20
2    DaNnY sMitH   76  Female         A-           Obesity        2022-09-22
3   andrEw waTtS   28  Female         O+          Diabetes        2020-11-18
4  adrIENNE bEll   43  Female        AB+            Cancer        2022-09-19

             Doctor                     Hospital Insurance Provider  \
0     Matthew Smith             Sons and Miller         Blue Cross
1   Samantha Davies                      Kim Inc           Medicare
2  Tiffany Mitchell                     Cook PLC              Aetna
3      Kevin Wells  Hernandez Rogers and Vang,           Medicare
4    Kathleen Hanna                 White-White              Aetna

   Billing Amount  Room Number Admission Type Discharge Date   Medication  \
0    18856.281306          328         Urgent     2024-02-02  Paracetamol
1    33643.327287          265      Emergency     2019-08-26    Ibuprofen
2    27955.096079          205      Emergency     2022-10-07      Aspirin
3    37909.782410          450       Elective     2020-12-18    Ibuprofen
4    14238.317814          458         Urgent     2022-10-09   Penicillin

   Test Results
0        Normal
1  Inconclusive
```

```
2        Normal
3      Abnormal
4      Abnormal
```

Descriptive Statistics

[ ]:
```python
# Select relevant numeric columns for analysis
numeric_columns = ['Age', 'Billing Amount']

# Perform Descriptive Statistics
print("Mean:\n", df[numeric_columns].mean())
print("\nMedian:\n", df[numeric_columns].median())
print("\nMode:\n", df[numeric_columns].mode().iloc[0])
print("\nStandard Deviation:\n", df[numeric_columns].std())
print("\nVariance:\n", df[numeric_columns].var())
print("\nRange:\n", df[numeric_columns].max() - df[numeric_columns].min())
print("\nSkewness:\n", df[numeric_columns].skew())
print("\nKurtosis:\n", df[numeric_columns].kurt())
```

```
Mean:
 Age                51.539459
Billing Amount    25539.316097
dtype: float64

Median:
 Age                52.000000
Billing Amount    25538.069376
dtype: float64

Mode:
 Age                38.000000
Billing Amount   -1316.618581
Name: 0, dtype: float64

Standard Deviation:
 Age                19.602454
Billing Amount    14211.454431
dtype: float64

Variance:
 Age                3.842562e+02
Billing Amount    2.019654e+08
dtype: float64

Range:
 Age                76.000000
Billing Amount    54772.768876
dtype: float64
```

```
Skewness:
 Age              -0.005735
Billing Amount    -0.000978
dtype: float64

Kurtosis:
 Age              -1.185576
Billing Amount    -1.190630
dtype: float64
```

Performing Inferential Statistics

[ ]:
```
# Inferential Statistics (One-sample t-test for Age)
age_values = df['Age']
population_mean = 40  # Hypothetical population mean for age
t_stat, p_value = stats.ttest_1samp(age_values, population_mean)
print(f"\nT-Statistic (Age): {t_stat}")
print(f"P-Value (Age): {p_value}")
```

```
T-Statistic (Age): 138.68245406411154
P-Value (Age): 0.0
```

Confidence Intervals

[ ]:
```
# Confidence Interval for Age
sample_mean_age = np.mean(age_values)
standard_error_age = stats.sem(age_values)
confidence_interval_age = stats.norm.interval(0.95, loc=sample_mean_age,␣
 ↪scale=standard_error_age)
print(f"95% Confidence Interval for Age: {confidence_interval_age}")
```

```
95% Confidence Interval for Age: (51.37637520273622, 51.702543716182696)
```

Regression Analysis

[ ]:
```
./../Billing Amount']
model = sm.OLS(y, X).fit()
print("\nRegression Summary (Billing Amount ~ Age):")
print(model.summary())
```

```
Regression Summary (Billing Amount ~ Age):
                            OLS Regression Results
==============================================================================
Dep. Variable:         Billing Amount   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.8149
Date:                Thu, 05 Sep 2024   Prob (F-statistic):              0.367
```

Time:	17:09:11	Log-Likelihood:	-6.0943e+05
No. Observations:	55500	AIC:	1.219e+06
Df Residuals:	55498	BIC:	1.219e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.568e+04	169.693	151.347	0.000	2.53e+04	2.6e+04
Age	-2.7781	3.077	-0.903	0.367	-8.810	3.254

Omnibus:	45130.804	Durbin-Watson:	1.985
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3278.455
Skew:	-0.001	Prob(JB):	0.00
Kurtosis:	1.809	Cond. No.	155.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.