Analysis of Children's Hospital Patient Data

Purpose of the Analysis

My Client-the hospital- wants to figure out how best to allocate resources and reduce costs. They have requested an analysis of past patient data to determine how the most common trends in demographic and diagnostic data affect hospital costs. The data was filtered, sorted, augmented with calculated columns, and statistically analysed for any findings that might be useful. This report contains both the analysis as well as Final Recommendations.

Methodology

I followed the standard Data analytics methodology.I was able to skip step one as the original data file (Excel) – which had been downloaded from Kaggle – was provided to me. It was stored locally in Excel format while I did the wrangling and data analysis.

Exploratory Data Analysis

In exploring the dataset, I discovered that there was data from 500 patients – with ages ranging from 0 (below a year) to 17.The data set also included Sex as a binary variable (female sex coded as 1), the length of hospital stay in days,the race of the patient coded numerically(1 to 6), the diagnosis code for the patients illness and the total hospital charges for each patient.To perform meaningful analysis for the Client, I would need to add additional columns of encoded dummy variables for the race attributes

Research Questions

- What age frequents the Hospital the most and what age has the maximum expenditure ?
- What diagnosis group has the most hospitalisation days and the highest expenditure?
- What are hospital costs by age and gender ?
- Can the length of stay be predicted from age, gender, and race since they are the most crucial factor for inpatients ?
- What variable mainly affects the hospital costs?

Data Wrangling

There was no need to do a lot of Data Wrangling of the data as it was provided in clean and complete excel spreadsheet format. However, I added some additional columns of “calculated” data in order to complete the analysis these columns were encoded dummy variables for the race attributes

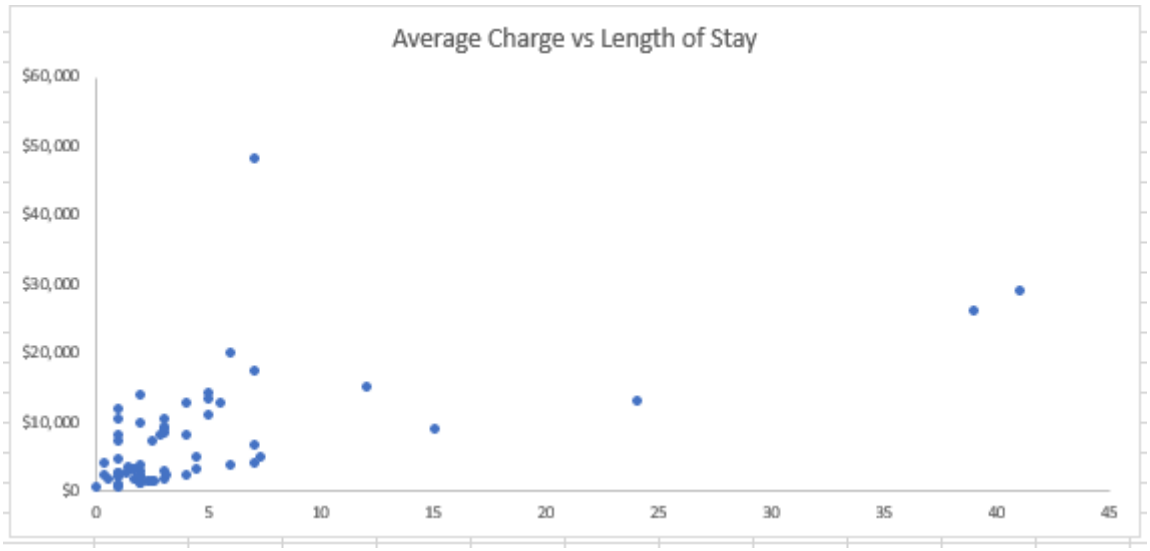
Data Analysis

For the Analysis the Excel Spreadsheet with the raw data was copied into a new spreadsheet called Working sheet Hospital costs.

- I inserted a pivot table on another sheet called Hospital costs pivot table. I added the Age variable in the rows, and count of age, sum of total charges and average total charges for each age in the columns. I then used conditional formatting to highlight the areas of interest and, we can see that children under a year accounted for over half of all patients in the dataset and they had the highest sum of total charges. However, on average patients aged 9,3,5, and 6 had the highest expenditure

Row Labels	Count of AGE	Sum of Total Charges	Average of Total Charges2
0	285	$629,657	$2,209
1	10	$37,744	$3,774
2	1	$7,298	$7,298
3	3	$30,550	$10,183
4	2	$15,992	$7,996
5	2	$18,507	$9,254
6	2	$17,928	$8,964
7	3	$10,087	$3,362
8	2	$4,741	$2,371
9	2	$21,147	$10,574
10	4	$24,469	$6,117
11	7	$13,621	$1,946
12	14	$53,602	$3,829
13	18	$31,135	$1,730
14	25	$64,643	$2,586
15	28	$108,673	$3,881
16	29	$69,149	$2,384
17	38	$174,777	$4,599
Grand Total	**475**	**$1,333,720**	**$2,808**

- I inserted another pivot table on the same Working sheet pivot table. I added the diagnosis codes in the rows, and count of length of stay, average length of stay,sum of total charges and average total charges for each diagnosis code in the columns. In order to visualise this better, I created a scatter plot of average charge versus length of stay.



We can see that most of the points are below 20,000 dollars charge and 5 days stay.We are looking for diagnosis codes with high expenditure and low length of stay - points in the upper left corner of the plot- while trying to avoid low expenditure and high length of stay diagnosis.

- This pivot table shows the sex and age variables in the rows, and sum of total charges and average total charges for each sex and age in the columns. I then used conditional formatting to highlight the areas of interest. We can see that on average male patients spent 500 dollars more than females. Females aged 3 and 9 had the highest average charge which is consistent with what was observed in the initial overall analysis while for males, ages 5 and 3 had the highest average charge.

Row Labels	Sum of Total Charges	Average of Total Charges2
0	**$699,472**	**$3,068**
0	$339,754	$2,178
1	$34,622	$4,328
2	$7,298	$7,298
3	$22,327	$11,164
4	$9,230	$9,230
5	$7,923	$7,923
6	$17,928	$8,964
7	$10,087	$3,362
8	$4,741	$2,371
9	$21,147	$10,574
10	$23,309	$7,770
11	$8,179	$1,636
12	$14,243	$2,849
13	$4,216	$1,054
14	$22,964	$5,741
15	$72,230	$7,223
16	$27,779	$4,630
17	$51,495	$3,961
1	**$634,248**	**$2,568**
0	$289,903	$2,247
1	$3,122	$1,561
3	$8,223	$8,223
4	$6,762	$6,762
5	$10,584	$10,584
10	$1,160	$1,160
11	$5,442	$2,721
12	$39,359	$4,373
13	$26,919	$1,923
14	$41,679	$1,985
15	$36,443	$2,025
16	$41,370	$1,799
17	$123,282	$4,931
Grand Total	**$1,333,720**	**$2,808**

- In order to find out if length of stay can be predicted from age, gender, and race, I created a table with the age and gender variable. I then used the IF function to encode dummy variables for the Race variable so I could run a regression analysis on it using the Excel data analysis pack. I ran the analysis on a new sheet called Hospital costs Regression Sheet.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.7589228	0.89943342	3.0674008	0.002278	0.99170955	4.526136	0.9917096	4.5261361
AGE	-0.038865	0.02238184	-1.736472	0.083108	-0.0828414	0.005111	-0.082841	0.0051106
FEMALE	0.3571474	0.31215992	1.1441167	0.253133	-0.2561867	0.970481	-0.256187	0.9704815
Race 1	0.0976284	0.89405505	0.1091973	0.913091	-1.6590174	1.854274	-1.659017	1.8542742
Race 2	0.0403314	1.38918255	0.0290325	0.97685	-2.6891445	2.769807	-2.689145	2.7698073
Race 3	-1.856551	3.37922682	-0.549401	0.58298	-8.4960805	4.782978	-8.496081	4.782978
Race 4	1.3941455	1.95334706	0.7137214	0.475739	-2.4438049	5.232096	-2.443805	5.232096
Race 5	-1.703232	2.0341584	-0.837316	0.402823	-5.6999616	2.293497	-5.699962	2.2934967
Race 6	-1.383342	2.3936779	-0.577915	0.563586	-6.0864581	3.319773	-6.086458	3.3197732

As seen above, the P-values for the variables do not show significance. I appears you cannot predict length of stay from a conglomerate of all these values

- Finally, I ran a regression analysis on all the variables to see if I could predict hospital costs length of stay, age and sex had significant p-values.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	776.17457	793.595267	0.9780484	0.328533	-783.09501	2335.444	-783.095	2335.4441
Length of Stay	740.7708	39.4427797	18.780897	**1.1E-59**	663.272946	818.2686	663.27295	818.26864
AGE	115.55651	19.6215599	5.8892621	**7.2E-09**	77.0037314	154.1093	77.003731	154.10929
FEMALE	-1008.989	273.189212	-3.693371	**0.00025**	-1545.7559	-472.2222	-1545.756	-472.2222
Race 1	-174.366	781.407892	-0.223143	0.823517	-1709.6896	1360.958	-1709.69	1360.9576
Race 2	745.24338	1214.13772	0.6138047	0.539629	-1640.3152	3130.802	-1640.315	3130.802
Race 3	-792.5794	2954.33033	-0.268277	0.788599	-6597.2982	5012.139	-6597.298	5012.1395
Race 4	267.95129	1708.09822	0.1568711	0.875411	-3088.1493	3624.052	-3088.149	3624.0519
Race 5	-809.7951	1779.11029	-0.455169	0.64919	-4305.4214	2685.831	-4305.421	2685.8313
Race 6	-781.2061	2092.7706	-0.373288	0.709096	-4893.1176	3330.705	-4893.118	3330.7055

Recap of Final Conclusions

- Children under a year old accounted for the most frequent patients, while children aged 9 and 3 had the highest average hospital costs both being over 10k dollars
- Patients with diagnosis code 911 had the highest expenditure paying about 48000 dollars on average for a staying an average of 7 days while those with code 609 had the longest average hospital stay of 41 days with an average cost of 29000 dollars
- On average male patients spent 500 dollars more than females. Females aged 3 and 9 had the highest average charge which is consistent with what was observed in the initial overall analysis while for males, ages 5 and 3 had the highest average charge
- You cannot predict length of stay from a conglomerate of the age, gender, and race variables, however it may be a different story if the variables are examined individually and i would recommend further analysis here
- Length of stay, age and sex could be used to predict hospital costs, there p-values were less than 0.05 and their co-efficients suggest that for each extra day stayed hospital costs increase by 740 dollars, for each year of age hospital costs increase by 115 dollars and finally, female patient costs are lower by about 1000 dollars