MARKETING 553Q Section C

Team 30 - Haocheng Chen, Xinyi Jiang, Kexin Mo, Jenny Park, Sachit Saraf

Predicting the Length of Stay of Each Patient

1. Introduction

Research Question: Predicting the Length of Stay of Each Patient

Business understanding

This project aims to build a model that predicts the length of stay of each patient with high accuracy so that hospitals can use the model to optimize resource allocation. Our findings from the project would be valuable to healthcare. At the beginning of the Covid19 Pandemic, we witnessed how unprepared the world's healthcare management was. It has raised alarms over the need to use various kinds of data.

For instance, analyzing patient information and the length of stay of a patient to improve the efficiency of a hospital. Suppose hospitals can identify the patients of the higher length of stay in advance at the time of admission. In that case, those patients can have their treatment optimized to minimize the length of stay for cost and lower the risk of staff and visitor infection.

Hypothesis

The key hypothesis is the length of each patient's stay could be related to factors like patient's, and hospital's conditions. Based on our data, we can improve healthcare management's efficiency in a hospital-like room and bed allocation planning.

Business Application

A healthcare management team can use the model we build to predict the number of rooms and beds required by entering necessary information ahead of time. It is not only cost-effective but also patient-friendly because higher efficiency will allow more patients in when needed. This model can be also helpful for new hospitals on how to structure the architecture, including how many rooms and beds are needed. It also helps existing hospitals to plan ahead for a future pandemic.

2. Data Understanding

Dataset source https://www.kaggle.com/nehaprabhavalkar/av-healthcare-analytics-ii

Variables

We obtained a fairly clean data from Kaggle. This dataset contains about 100,000 patient records from hospitals in various locations. Each row in the dataset represents a session of a patient's stay. From our observation, we found variables related to hospitals' information like locations and bed grades will affect the length of stay of each patient. (See Appendix 1.1)

To predict the patient length of stay, we would classify patients by segmentation using the key independent variables as shown in Appendix 1.2.

Data cleaning

The original dataset contains more than 300,000 rows. It's challenging to run an analysis with this tremendous data. Hence, we decided to use R to remove 50% of the records from the original data randomly and dropped all records with duplicates and null values. Then, we randomly split 80% of the dataset as train data, and 20% as test data. In our analysis, we utilized the test data to check for the overall accuracy. At the end of our data cleaning process, we checked for the dataset's basic validation of the cleaned data frame.

3. Exploratory data analysis

We firstly looked at the distribution of the dependent variable. (See Appendix 3.1)The Stay variable has 159,162 observations and 11 values. Its distribution is skewed to the right. "21-30" value has the highest frequency and accounts for 27.42% of the observations. "61-70" value has the least frequency and accounts for 0.83% of the data. About half of the patients stay at the hospital for about one month or less. Few patients, about 8%, stay at the hospital for more than two months.

There are three numerical variables in our dataset - admission_deposit, available_extra_rooms, and visitors_with_patient. We calculated the mean, range and standard deviation of the three variables. (See Appendix 3.2)

We also calculated the descriptive statistics for categorical independent variables. For one instance, in terms of Age, there are 10 categories. The patients aged between 31-40 account for the largest proportion - 20.02%. And most of the patients, about 55%, are middle-aged. (See Appendix 3.3)

Then we investigated the relationship between the dependent and independent variables. (See Appendix 3.4) We assumed that the hospital attributes, like bed grade or location, might have an effect on the stay length. We plotted the relationship between bed grade and stay length. We found that most patients are

attributed to bed-grade 2 and 3. For the Department variable, we know that most of the patients belong to the Gynecology, Anesthesia, and Radiotherapy department. We also drew heat maps to show the relationships among three variables (hospital_region_code & hospital_type_code & stay, available_extra_room & ward_type & stay). However, the graphs could not indicate any significant relationship among these variables. So we planned to intuitively select some variables related to hospitals to build our Model 1, testing whether hospital attributes affect the patients' stay length or not.

Then we assumed that patients' personal attributes affected their stay length. We selected four variables that are important to our dependent variable Stay length: Severity of illness, Type of admission, Age, and Visitors with patient. (See Appendix 3.5) From the histogram based on Type of admission and Stay, most patients are attributed to Trauma. In terms of severity of illness, most patients are attributed to Moderate illness. From the heat map, we found that if the patient was older and had more visitors, they were more likely to stay longer in hospital. From these graphs, we concluded that the patients' personal attributes might have more significant relationships with the stay length. Therefore, we planned to add variables related to patients in our Model 2 to improve our prediction accuracy.

4. Modeling

We performed two ordered logistic models with different numbers of independent variables because the response variable, stay, is a categorical feature with hierarchy and not a continuous numeric feature. The models generated the predicted category of stay and we calculated the percentage of true positive rate, accuracy, as the evaluation metric.

The first model we created only included variables that are related to hospitals. The dependent variable is the time of staying for the patients, and the 5 independent variables are bed_grade, city_code_hospital, department, ward_type, and available_extra_room. The testing global hypotheses are significant for all tests, and all variables are significant under type 3 analysis. The accuracy of model one is 0.2762. We believe this model would have a better prediction than guessing stays of patients randomly because the accuracy is greater than the baseline 0.09.

To determine if the accuracy of the existing model would improve with patient's information, we added additional 6 variables to the first model: age,

severity_of_illness,type_of_admission, admission_deposit, available_extra_room, and visitors_with_patien. The accuracy of the second model is 0.35554, which showed a better prediction than the first model.

5. Results and managerial takeaways

To better understand each variable from our models, we utilized the analysis of maximum likelihood estimation to determine values for the parameters of our model. In this estimation, the categorical variables would break into individual parts to compare with a fixed variable. In the output (appendix 2.6), we found the log odds of stay increases for bed grade 1 and decreases for 2 and 3 compared to bed grade 4. From this result, we think bed grade 1 has the most severity of illness of patients among the others. The log odds of stay decreased by 0.5243 for city code hospital 7 compared to city code hospital 13. These statistics showed the medical facility could be more advanced in code hospital 7 than 13. Furthermore, the log odds of stay in radiotherapy are not that different from the log odds of stay in surgery. The reason for the correlation might be radiotherapy and surgery are treatments often used for similar diseases as cancer to remove shrink tumors. For variable ward type, we found that only S is significantly different from U. A possible reason for this result might be that S and U represent two distinct departments like children and elders. Lastly, we found one unit increase in extra room decreases log odds of stay by 0.0815 which also showed a logical change in the accuracy of this model.

Some of the findings for model 2 were like model 1, but with the addition of patient information, we could clearly interpret how the condition of each patient impacted the patient's stay (appendix 2.7). For a patient in the age group of 0-10, the log odds of stay decreased by 0.4813 when compared to a patient in the age group of 91-100. However, an interesting finding was that a patient in the age group of 21-30 did not have a significant difference in log odds of stay with a patient in the age group of 91-100. Interpretation of bed grade in model 2 differed from that of model 1. Log odds of stay increases for bed grades of 1 and 2 and decreases for 3 when compared to 4. City code of a hospital showed that hospitals from city codes 4 and 10 are not significantly different from city code 13. We can assume that these cities may be adjacent to one another or that these cities share similar traits and that hospitals in these regions have similar levels of facilities.

From model 2, we could also see the impact of the type of admission to stay. We see that patients who came to the hospital for trauma have a higher probability to stay longer than patients who were admitted for emergency or urgent. This finding makes sense because trauma patients are usually ones who have a potentially life-threatening physical injury. Some additional information we see from model 2 also includes admission deposits and the number of visitors. When deposit increases by $1000, log odds of stay increase by 0.071 and one additional visitor increases log odds of stay by 0.756.

While both models showed higher accuracy than the baseline of 0.09, model 2 showed higher accuracy than model 1. We predicted that introducing patient information to the model and, in fact, our accuracy proved that our prediction was correct. This prediction is logical because the stay of a patient heavily depends on the condition of the patient. We see that a lower age has a lesser probability to stay longer. The severity of illness of a patient also provides a great impact in predicting the probability of stay. A patient with an extreme condition showed a higher probability of stay and a patient with a minor condition showed a lower probability of stay. Lastly, through trying different variable variations, we discovered that the number of visitors increased the accuracy greatly. While the number of visitors serves as an important feature in the prediction, it is not an informative variable because we cannot obtain visitor information at the time patients check-in.

6. Limitations

There's not enough specific information about patients in the database. Most of them are information about hospitals' conditions, ward conditions. Since the original purpose is to predict patients' stay, we could also include more basic information like gender, marital status of patients, instead of just ages.

Since one of our goals is trying to target and improve hospitals' efficiency due to the Covid-19 pandemic, we should also put focus on Covid-19 patients as well. Including columns like:" Whether patients have Covid-19", "Whether patients had Covid-19 before", etc could be more information for us during the special period like Covid-19 Pandemic. Then it's more meaningful for hospitals to predict patients' stay lengths and deal with such emergency situations.

Appendix

2.1 Variables related to hospitals' information

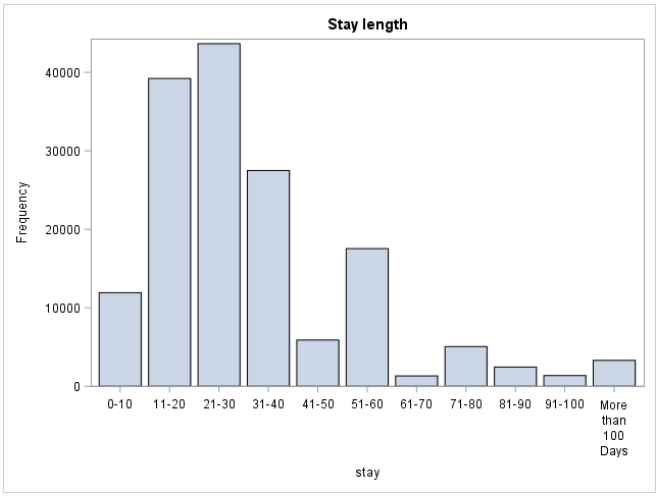
Hospital_code	Unique code for the Hospital
Hospital_type_code	Unique code for the type of Hospital
City_Code_Hospital	City Code of the Hospital
Hospital_region_code	Region Code of the Hospital
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital
Department	Department overlooking the case (gynecology, anesthesia,radiotherapy, surgery, TB & chest disease)
Ward_Type	Code for the Ward type (P, Q, R, S, T, U)
Ward_Facility_Code	Code for the Ward Facility (A, B, C, D, E, F)
Bed Grade	Condition of Bed in the Ward (1.0-4.0)

2.2 Variables related to patients' information

Age	Age range of patient
Severity of Illness	Severity of the illness recorded at the time of admission
Visitors with Patient	Number of Visitors with the patient
Type of Admission	Classify the patient's visiting type (emergency, trauma, or urgent)
City_Code_Patient	Code for identifying which city the patient from
Stay	Length of stay for the patient

3.1 Distribution of Stay variable

stay	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-10	11918	7.49	11918	7.49
11-20	39212	24.64	51130	32.12
21-30	43649	27.42	94779	59.55
31-40	27488	17.27	122267	76.82
41-50	5886	3.70	128153	80.52
51-60	17539	11.02	145692	91.54
61-70	1315	0.83	147007	92.36
71-80	5054	3.18	152061	95.54
81-90	2444	1.54	154505	97.07
91-100	1360	0.85	155865	97.93
More than 100 Days	3297	2.07	159162	100.00



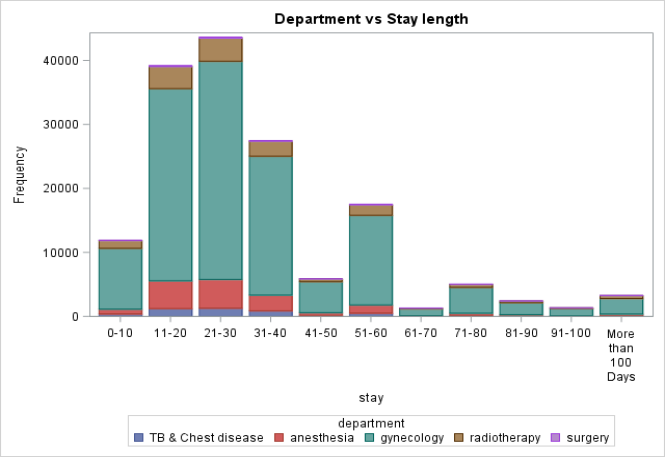
3.2 Descriptive statistics about numerical variables

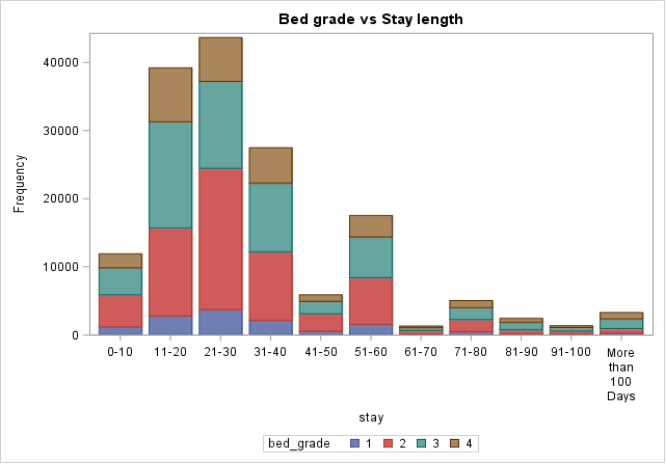
Variable	N	Mean	Std Dev	Minimum	Maximum
admission_deposit	159162	4884.70	1089.54	1809.00	11008.00
available_extra_rooms	159162	3.1980058	1.1735371	0	21.0000000
visitors_with_patient	159162	3.2821151	1.7690786	0	32.0000000

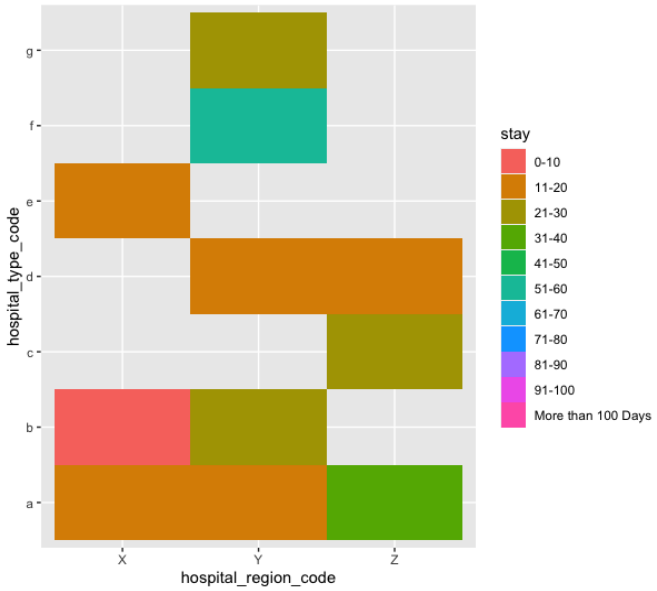
3.3 descriptive statistics for Age variable

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-10	3206	2.01	3206	2.01
11-20	8359	5.25	11565	7.27
21-30	20581	12.93	32146	20.20
31-40	31870	20.02	64016	40.22
41-50	31598	19.85	95614	60.07
51-60	24180	15.19	119794	75.27
61-70	16829	10.57	136623	85.84
71-80	17840	11.21	154463	97.05
81-90	4050	2.54	158513	99.59
91-10	649	0.41	159162	100.00

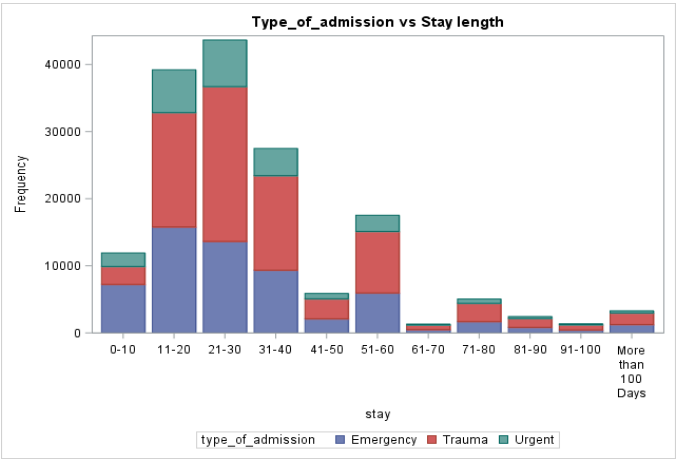
3.4 The relationship between stay length and different independent variables related to hospitals

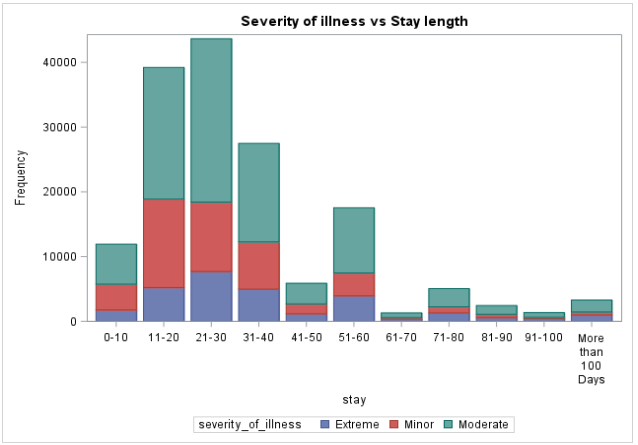


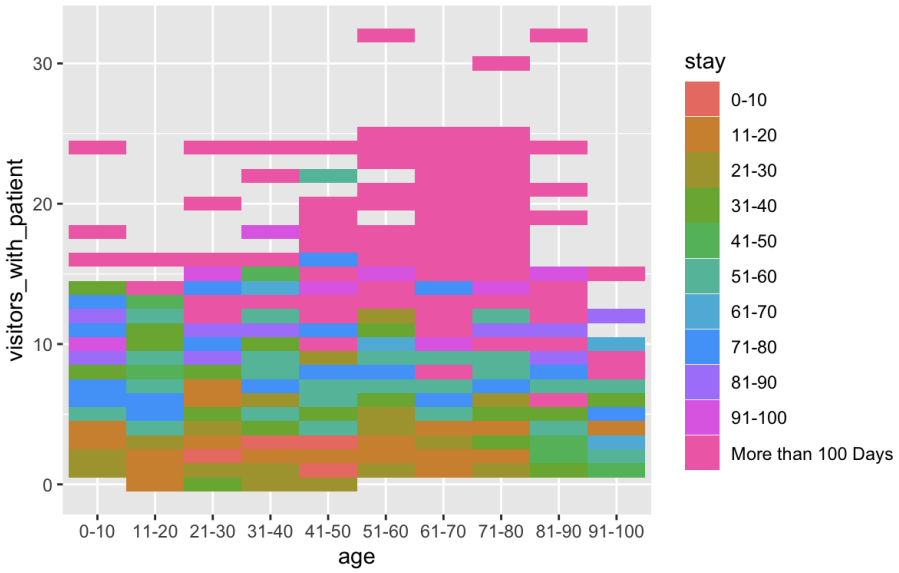




3.5 The relationship between patient-related variables and Stay length







3.6 Maximum Likelihood Estimates - Model 1

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**More than 100 Days**	1	-3.7910	0.2147	311.8647	<.0001
Intercept	**91-100**	1	-3.4349	0.2144	256.6412	<.0001
Intercept	**81-90**	1	-2.9900	0.2142	194.8655	<.0001
Intercept	**71-80**	1	-2.4141	0.2140	127.2241	<.0001
Intercept	**61-70**	1	-2.2979	0.2140	115.3023	<.0001
Intercept	**51-60**	1	-1.3057	0.2139	37.2663	<.0001
Intercept	**41-50**	1	-1.0767	0.2139	25.3454	<.0001
Intercept	**31-40**	1	-0.2348	0.2138	1.2059	0.2722

Intercept	**21-30**	1	0.9429	0.2139	19.4393	<.0001
Intercept	**11-20**	1	2.7384	0.2141	163.6692	<.0001
bed_grade	**1**	1	0.0970	0.0138	49.2766	<.0001
bed_grade	**2**	1	-0.0182	0.00825	4.8431	0.0278
bed_grade	**3**	1	-0.0674	0.00852	62.4947	<.0001
city_code_hospital	**1**	1	-0.1204	0.0131	84.8169	<.0001
city_code_hospital	**2**	1	0.2181	0.0136	256.1726	<.0001
city_code_hospital	**3**	1	-0.0358	0.0160	4.9863	0.0255
city_code_hospital	**4**	1	-0.0797	0.0228	12.2363	0.0005
city_code_hospital	**5**	1	-0.0673	0.0163	17.0525	<.0001
city_code_hospital	**6**	1	0.1937	0.0143	183.5388	<.0001
city_code_hospital	**7**	1	-0.5243	0.0157	1108.8023	<.0001
city_code_hospital	**9**	1	-0.1815	0.0178	104.4026	<.0001
city_code_hospital	**10**	1	0.1136	0.0353	10.3456	0.0013
city_code_hospital	**11**	1	0.0668	0.0215	9.6344	0.0019
department	**TB & Chest disease**	1	-0.1368	0.0281	23.7200	<.0001
department	**anesthesia**	1	-0.2444	0.0216	128.0820	<.0001
department	**gynecology**	1	0.0495	0.0179	7.6797	0.0056
department	**radiotherapy**	1	0.00156	0.0215	0.0053	0.9421
ward_type	**P**	1	-0.0280	0.2153	0.0169	0.8965
ward_type	**Q**	1	-0.1513	0.2127	0.5062	0.4768

ward_type	**R**	1	0.0833	0.2127	0.1534	0.6953
ward_type	**S**	1	0.7146	0.2128	11.2813	0.0008
ward_type	**T**	1	0.1797	0.2213	0.6595	0.4167
available_extra_room		1	-0.0815	0.00482	285.8742	<.0001

3.7 Maximum Likelihood Estimates - Model 2

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**More than 100 Days**	1	-7.6301	0.2199	1203.4722	<.0001
Intercept	**91-100**	1	-7.1351	0.2194	1057.2091	<.0001
Intercept	**81-90**	1	-6.5152	0.2190	885.0697	<.0001
Intercept	**71-80**	1	-5.7109	0.2186	682.4279	<.0001
Intercept	**61-70**	1	-5.5536	0.2186	645.7000	<.0001
Intercept	**51-60**	1	-4.1925	0.2182	369.1496	<.0001
Intercept	**41-50**	1	-3.8803	0.2182	316.3625	<.0001
Intercept	**31-40**	1	-2.7729	0.2180	161.7459	<.0001
Intercept	**21-30**	1	-1.3562	0.2180	38.7154	<.0001
Intercept	**11-20**	1	0.5847	0.2181	7.1890	0.0073
age	**0-10**	1	-0.4813	0.0335	206.8162	<.0001
age	**11-20**	1	-0.2699	0.0223	146.4739	<.0001
age	**21-30**	1	-0.00423	0.0162	0.0680	0.7943

age	**31-40**	1	0.0888	0.0144	37.9807	<.0001
age	**41-50**	1	-0.0388	0.0143	7.3237	0.0068
age	**51-60**	1	-0.1764	0.0154	131.3110	<.0001
age	**61-70**	1	-0.1628	0.0172	89.4419	<.0001
age	**71-80**	1	0.1331	0.0168	62.3627	<.0001
age	**81-90**	1	0.4750	0.0299	253.0299	<.0001
bed_grade	**1**	1	0.0956	0.0147	42.1045	<.0001
bed_grade	**2**	1	0.1305	0.00859	231.1774	<.0001
bed_grade	**3**	1	-0.0900	0.00899	100.2099	<.0001
city_code_hospital	**1**	1	-0.0541	0.0133	16.6180	<.0001
city_code_hospital	**2**	1	0.3161	0.0138	521.1345	<.0001
city_code_hospital	**3**	1	0.0669	0.0163	16.8878	<.0001
city_code_hospital	**4**	1	0.00687	0.0231	0.0884	0.7662
city_code_hospital	**5**	1	-0.0610	0.0165	13.6195	0.0002
city_code_hospital	**6**	1	0.1208	0.0146	68.6265	<.0001
city_code_hospital	**7**	1	-0.5202	0.0164	1003.7399	<.0001
city_code_hospital	**9**	1	-0.1657	0.0180	84.7827	<.0001
city_code_hospital	**10**	1	0.0116	0.0358	0.1052	0.7457
city_code_hospital	**11**	1	-0.0421	0.0219	3.7139	0.0540
department	**TB & Chest disease**	1	-0.1914	0.0285	44.9879	<.0001
department	**anesthesia**	1	-0.2277	0.0221	106.2136	<.0001

department	**gynecology**	1	0.1104	0.0182	36.6765	<.0001
department	**radiotherapy**	1	-0.0193	0.0219	0.7764	0.3782
severity_of_illness	**Extreme**	1	0.3005	0.00950	1001.3872	<.0001
severity_of_illness	**Minor**	1	-0.3245	0.00862	1417.1729	<.0001
type_of_admission	**Emergency**	1	-0.3316	0.00792	1754.2659	<.0001
type_of_admission	**Trauma**	1	0.2802	0.00728	1480.8121	<.0001
ward_type	**P**	1	-0.9119	0.2174	17.5893	<.0001
ward_type	**Q**	1	-0.2901	0.2147	1.8263	0.1766
ward_type	**R**	1	0.2203	0.2147	1.0529	0.3048
ward_type	**S**	1	0.8593	0.2148	16.0119	<.0001
ward_type	**T**	1	0.4168	0.2234	3.4810	0.0621
admission_deposit		1	0.000071	4.854E-6	211.1888	<.0001
available_extra_room		1	-0.1850	0.00509	1322.0495	<.0001
visitors_with_patien		1	0.7560	0.00397	36230.2450	<.0001