Inpatient Analysis & Predicting Length of StayData Data Science Project

Project outcomes or summery

Hospital Inpatient Discharges (SPARCS De-Identified): 2015

Data obtained from https://www.kaggle.com/datasets/jonasalmeida/2015-deidentified-ny-inpatient-discharge-sparcs/data

About Dataset:

Public Health Data

This is the public dataset made available at https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8 by the Dept of Health of New York state. The following description can be found at that page:

* The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. The de-identified data file does not contain data that is protected health information (PHI) under HIPAA. The health information is not individually identifiable; all data elements considered identifiable have been redacted. For example, the direct identifiers regarding a date have the day and month portion of the date removed.

* The aim of this project is to conduct inpatient analysis and predict the length of stay in the hospital using the parameters likely to be available when the patient is admitted.

* The data is unclean, has missing values, and contains 2.35 million rows and 37 columns. It may not be necessary to include all instances and features to achieve the goal of this project.

Step1: Data Cleaning & preparation Summary:

- Modifying columns names to be more suitable.

- As the data contains 2.35 million rows and 37 columns then for any columns that have more than 20% missing values, we drop the columns. Affected columns as follows:

1. payment_typology_2 by 32.490264 %
2. payment_typology_3 by 70.124635 %
3. operating_provider_license_number by 26.114227 %
4. other_provider_license_number by 96.957239 %.

- For less than 20 % we drop rows. affected columns as follows:

1. health_service_area by 0.124166 %
2. hospital_county by 0.124166 %
3. operating_certificate_number by 0.124166 %
4. facility_id by 0.124166 %
5. zip_code__3_digits by 0.188744 %
6. apr_severity_of_illness_description by 0.004777 %
7. apr_risk_of_mortality by 0.004777 %
8. attending_provider_license_number by 0.124166 %

- After checking info, unique, and value counts we found that:

1. zip_code__3_digits feature has strange value 'OOS' repeated 67135 times which refer to Out of State zip codes replaced it with 005 to indicate out of state. Also, changing dtype to numeric.
2. gender feature has strange value 'U' repeated 39 times these values raws dropped.
3. length_of_stay feature has '120 +', we create new feature for over_120_stay then I treated all entries '120 +' in length of stay as 120. also dtype changed to numerical.discharge_year feature has only 2015 value. it will not be useful so drop it.
4. attending_provider_license_number feature it is not useful and we drop it.
5. abortion_edit_indicator feature has only 'N' value. it will not be useful we drop it.
6. total_charges feature has Dollar signs. Sign removed and replaced dtype to numerical.
7. total_costs feature has Dollar signs. Sign removed and replaced dtype to numerical.

- Finding 10,598 duplicated record the dealing with them done by dropping them and reset the index.
- Saving file as cleaned_data.csv after cleaning process for next Analysis steps.

Step2: Analysis outcomes or summary:

The main objective of this analysis is:

1. Statistical analysis: for the important numerical and categorical features.

2. The Correlation & Correlation heatmap: for important numerical feature.

3. Analysis and visualization around the following quotations:

1. What is the patient distribution of most important features of New York State during the 2015-year dataset?
2. Calculate average Length of Stay Ratio for top or all if applicable features (APR Severity of Illness Code, APR Risk of Mortality, APR Medical Surgical Description, Payment Typology 1, Emergency Department Indicator, APR MDC Description, Age Group, Gender).
3. What is the relationship between Birth Weight and length of stay?

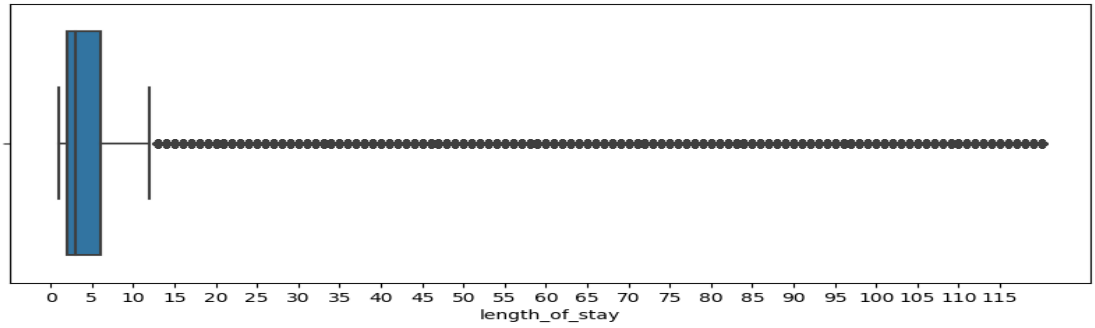
	length_of_stay	total_charges	total_costs
count	2331584	2331584	2331584
mean	5.5	43393.74	16050.24
std	8.05	80606.97	32455.39
min	1	0.01	0
25%	2	12145.07	4764.31
50%	3	23634	8841.34
75%	6	46819.13	16905.13
max	120	7248390.82	5236614.76

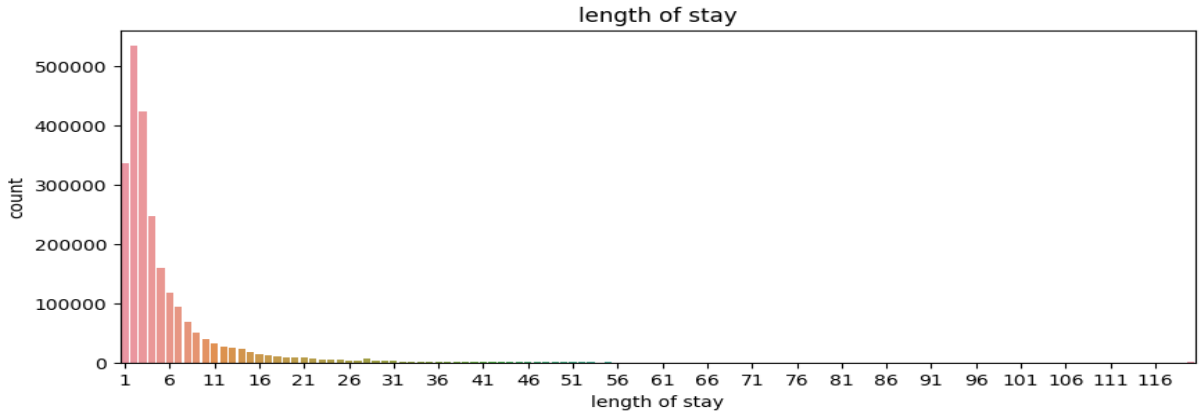
1. Statistical analysis outcomes: for the important numerical and categorical features.

- **First Statistical Description for Numeric Features for the State of New York 2015 Inpatient Discharges:** We are going to focus on important numerical features:
    1. **Length_of_Stay Feature:** Ranges from 1 to 120 days. The mean length of stay is almost 6 days for the entire population, and the median is 3 days. As the mean is greater than the median, it reflects that the data is right-skewed, and there are outliers in the upper values.
    2. **Total_Charges:** Ranges from 0.01 to 7,248,390.82 dollars. The mean is 43,393.74 dollars, and the median is 23,634 dollars. As the mean is greater than the median, it reflects that the data is right-skewed, and there are outliers in the upper values.
    3. **Total_Costs Feature:** Ranges from 0.00 to 5,236,614.76 dollars. The mean is 16,050.24 dollars, and the median is 8,841.34 dollars. As the mean is greater than the median, it reflects that the data is right-skewed, and there are outliers in the upper values.

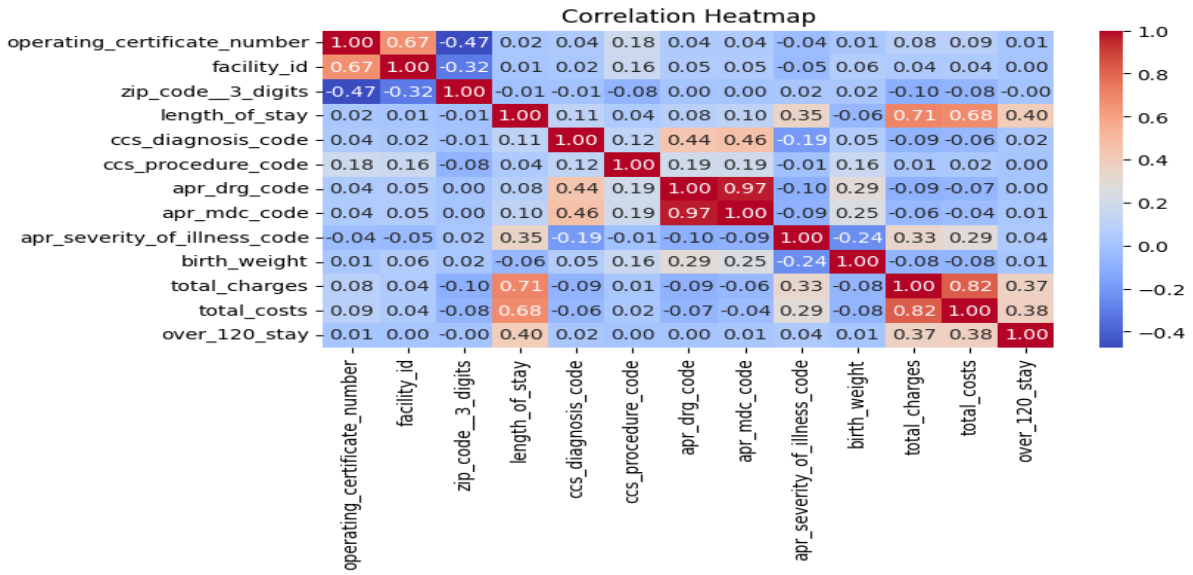
- **Second Statistical Description for Categorical Features:**
    1. **Health_Service_Area Feature:** Out of **8** areas, the most frequented area is **New York City**, which received **1,083,178** patients out of **2,331,584** total patients.
    2. **Hospital_County Feature:** Out of **57** counties, the most frequented county is **Manhattan**, which received **397,939** patients out of the total.
    3. **Facility_Name Feature:** Out of **214** facilities, the most frequented facility is **Mount Sinai Hospital**, which received **55,005** patients out of the total.
    4. **Age_Group Feature:** Out of **5** age groups, the most frequented age group is **50 to 69** years old, with **644,707** patients out of the total.
    5. **Gender Feature:** Out of **2** genders, the most frequented is **Female**, with **1,297,197** female patients out of the total.
    6. **Race Feature:** Out of **4** races, the most frequented is **White**, with **1,328,630** white race patients out of the total.
    7. **Ethnicity Feature:** Out of **4** ethnicities, the most frequented is **Not Hispanic/Spanish**, with **1,328,630** patients out of the total.
    8. **Type_of_Admission Feature:** Out of **6** admission types, the most frequented is admission from **Emergency**, with **1,484,292** patients out of the total.
    9. **Patient_Disposition Feature:** Out of **19** disposition types, the most frequented is **Home or Self Care**, with **1,557,762** patients out of the total.

10. **CCS_Diagnosis_Description Feature:** Out of **263** diagnosis descriptions, the most frequented is **Liveborn**, with **217,052** patients out of the total.
11. **CCS_Procedure_Description Feature:** Out of **232** procedure descriptions, the most frequented is **NO PROC**, with **607,287** patients out of the total.
12. **APR_DRG_Description Feature:** Out of **314** descriptions, the most frequented is **"Neonate birthweight >2499g, normal newborn or neonate..."** with **188,536** patients out of the total.
13. **APR_MDC_Description Feature:** Out of **25** descriptions, the most frequented is **"Diseases and Disorders of the Circulatory System"** with **289,397** patients out of the total.
14. **APR_Severity_of_Illness_Description Feature:** Out of **4** descriptions, the most frequented is **Moderate**, with **895,145** patients out of the total.
15. **APR_Risk_of_Mortality Feature:** Out of **4** descriptions, the most frequented is **Minor**, with **1,375,309** patients out of the total.
16. **APR_Medical_Surgical_Description Feature:** Out of **2** descriptions, the most frequented is **Medical**, with **1,766,548** patients out of the total.
17. **Payment_Typology_1 Feature:** Out of **10** descriptions, the most frequented is **Medicare**, with **875,749** patients out of the total.
18. **Emergency_Department_Indicator Feature:** Out of **2** descriptions, the most frequented is **Y**, with **1,365,513** patients out of the total.





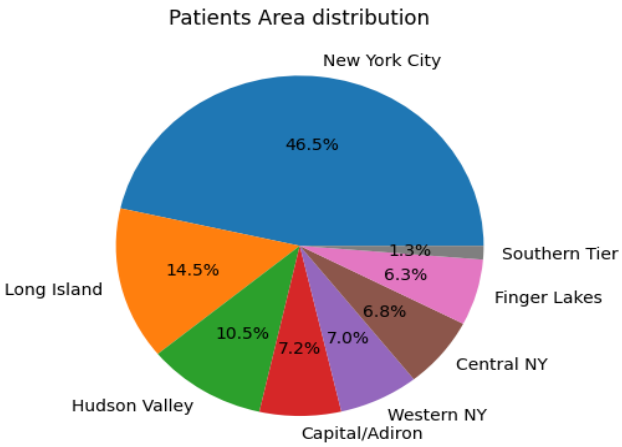
2. The Correlation & Correlation heatmap outcomes: for important numerical feature.



Although the correlation matrix may not provide meaningful insights for all columns, it is evident that the APR Severity of Illness Code shows a strong positive correlation with the length of stay, as do total charges and total costs. Additionally, CCS Diagnosis code appears to have a slight positive correlation with the length of stay. Positive correlations can also be observed in the dataset between features such as CCS Diagnosis codes and APR DRG codes.

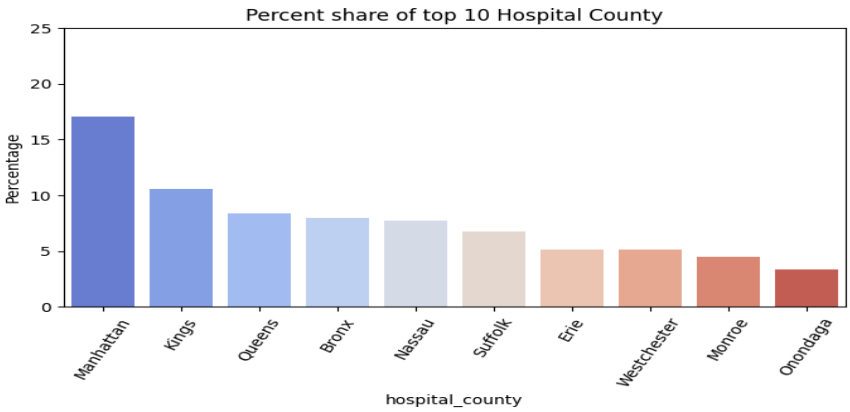
3. Analysis and Visualization Outcomes:

1. What is the distribution percentage of patients among cities of New York State during the year 2015?
    - New York City: 46.46%
    - Long Island: 14.50%
    - Hudson Valley: 10.51%
    - Capital/Adirondack: 7.17%
    - Western NY: 7.02%
    - Central NY: 6.77%
    - Finger Lakes: 6.28%
    - Southern Tier: 1.29%

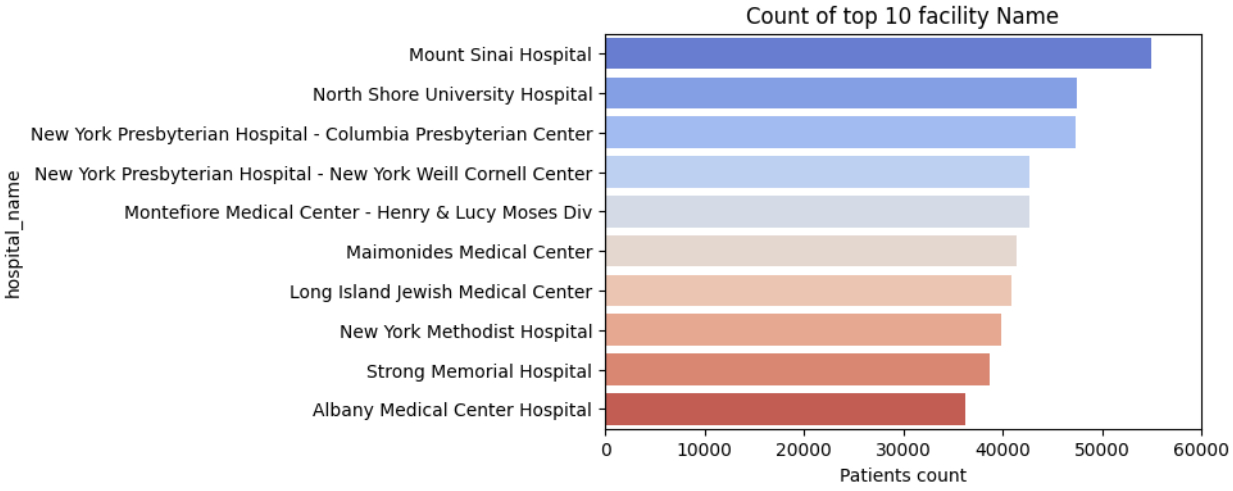


 Project prepared by: Diaa Aldein Al Sayed Ibrahim Osman

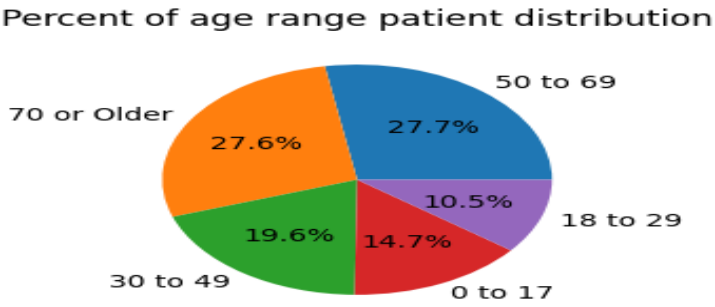
2. What is the percent share of the top 10 Hospital Counties?
    - Manhattan: 17.07%
    - Kings: 10.58%
    - Queens: 8.38%
    - Bronx: 7.99%
    - Nassau: 7.75%
    - Suffolk: 6.75%
    - Erie: 5.16%
    - Westchester: 5.11%
    - Monroe: 4.51%
    - Onondaga: 3.35%



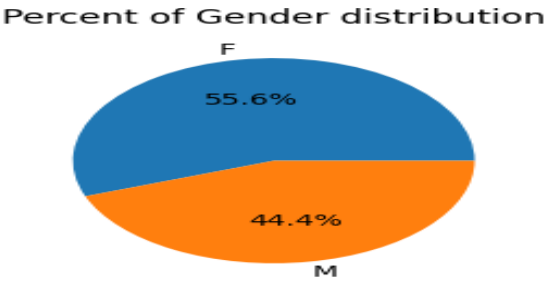
3. What are the top 10 hospitals that received patients during the year 2015?
    - Mount Sinai Hospital: 55005
    - North Shore University Hospital: 47463
    - New York Presbyterian Hospital - Columbia Presbyterian Center: 47399
    - New York Presbyterian Hospital - New York Weill Cornell Center: 42724
    - Montefiore Medical Center - Henry & Lucy Moses Div: 42715
    - Maimonides Medical Center: 41466
    - Long Island Jewish Medical Center: 40850
    - New York Methodist Hospital: 39925
    - Strong Memorial Hospital: 38653
    - Albany Medical Center Hospital: 36289



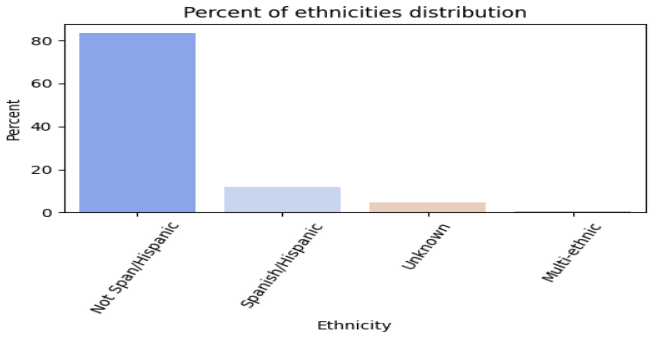
4. What is the distribution percentage of age groups among patients during the year 2015?
   - 50 to 69: 27.65%
   - 70 or Older: 27.58%
   - 30 to 49: 19.56%
   - 0 to 17: 14.68%
   - 18 to 29: 10.53%



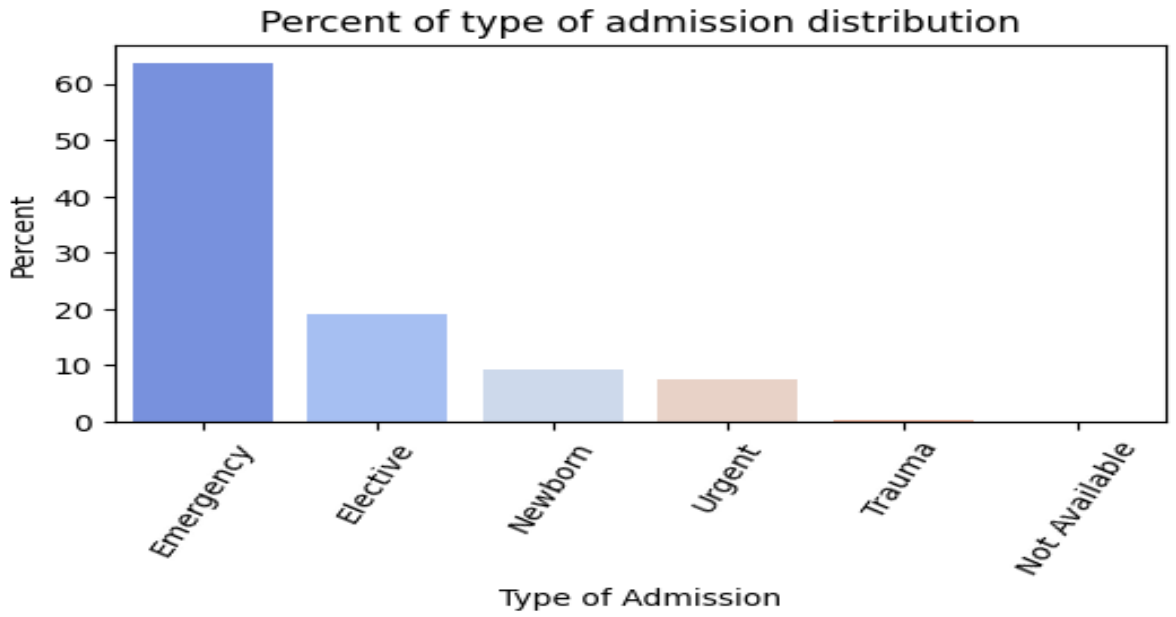
5. What is the percentage of gender distribution among patients?
   - Female: 55.64%
   - Male: 44.36%



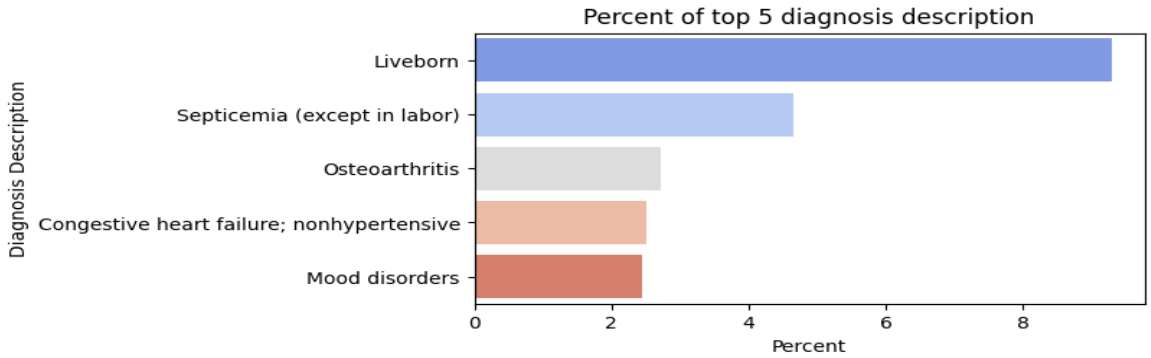
6. What is the distribution percentage of ethnicities among patients?
   - Not Spanish/Hispanic: 83.33%
   - Spanish/Hispanic: 11.83%
   - Unknown: 4.47%
   - Multi-ethnic: 0.37%



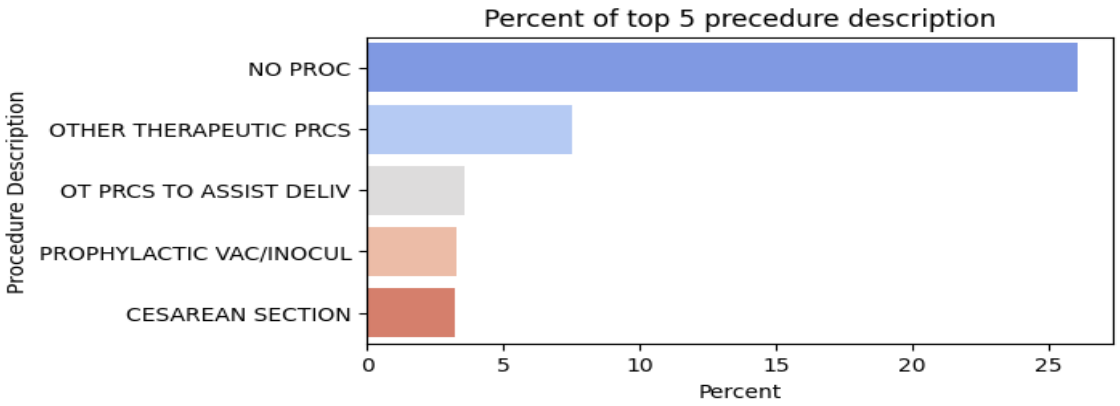
7. What is the distribution percentage of the type of admission among patients?
    - Emergency: 63.66%
    - Elective: 19.14%
    - Newborn: 9.32%
    - Urgent: 7.56%
    - Trauma: 0.27%
    - Not Available: 0.05%



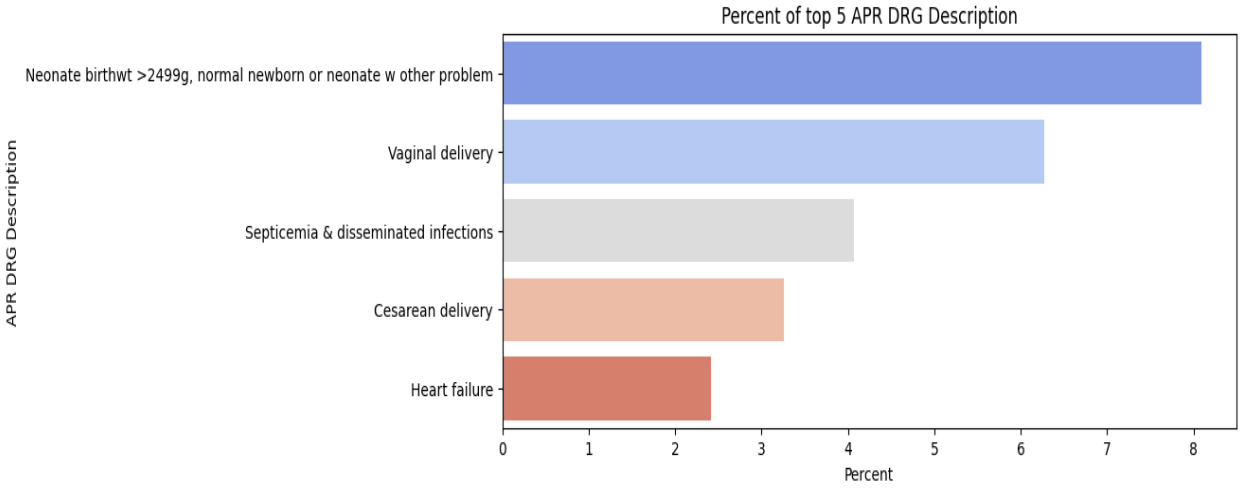
8. What are the top 5 diagnosis descriptions among patients, and what percentage does each one represent?
    - Liveborn: 9.31%
    - Septicemia (except in labor): 4.66%
    - Osteoarthritis: 2.72%
    - Congestive heart failure; nonhypertensive: 2.52%
    - Mood disorders: 2.44%



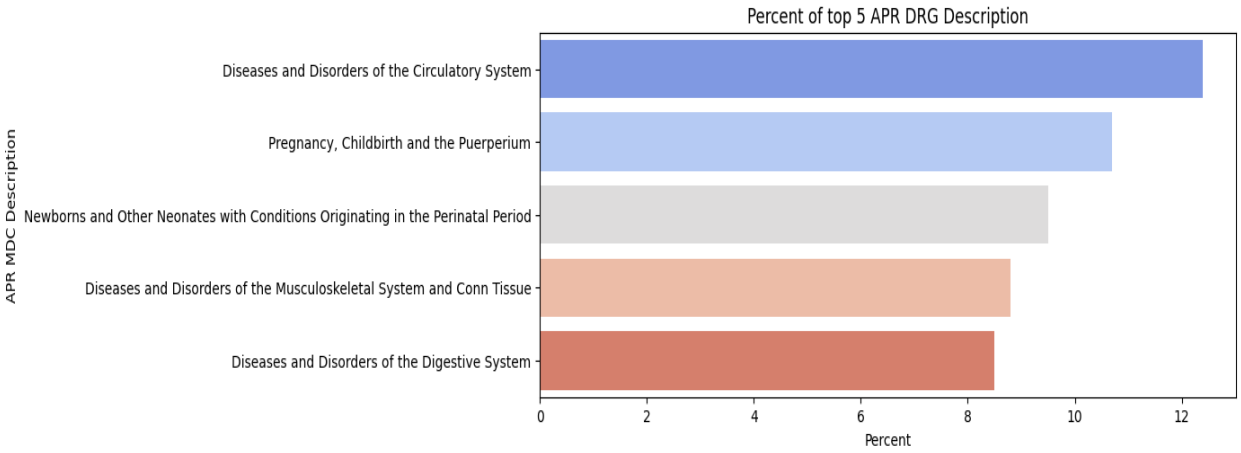
9. What are the top 5 procedure descriptions among patients, and what percentage does each one represents?
    - NO PROC: 26.05%
    - OTHER THERAPEUTIC PRCS: 7.51%
    - OT PRCS TO ASSIST DELIV: 3.56%
    - PROPHYLACTIC VAC/INOCUL: 3.28%
    - CESAREAN SECTION: 3.20%



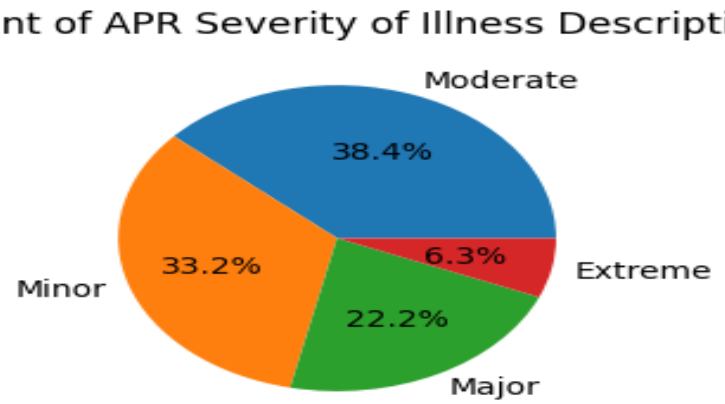
10. What are the top 5 apr_drg_descriptions among patients, and what percentage does each one represents?
    - Neonate birth weight >2499g, normal newborn or neonate w other problem: 8.09%
    - Vaginal delivery: 6.27%
    - Septicemia & disseminated infections: 4.07%
    - Cesarean delivery: 3.25%
    - Heart failure: 2.42%



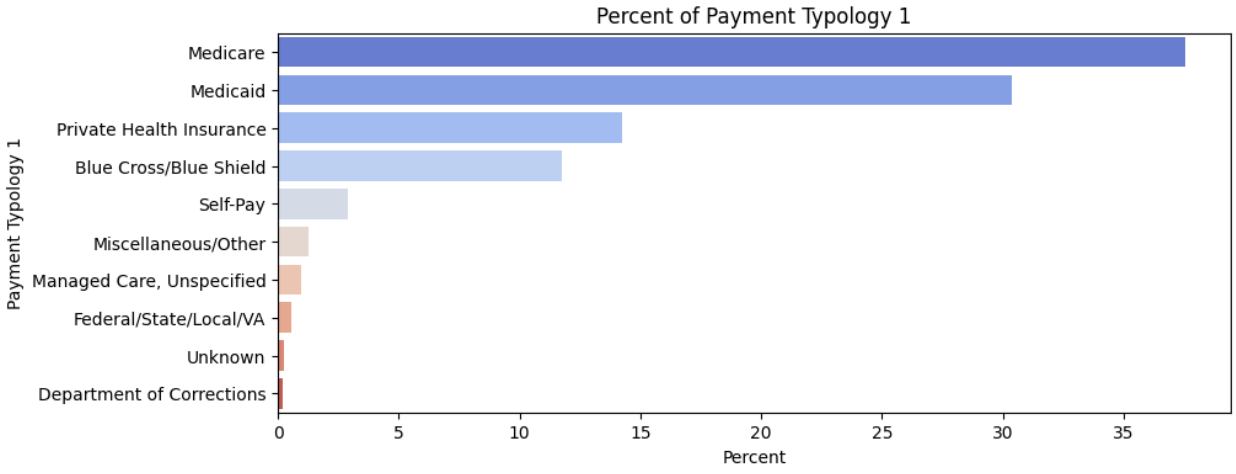
11. What are the top 5 apr_mdc_descriptions among patients, and what percentage does each one represent?
    - Diseases and Disorders of the Circulatory System: 12.4%
    - Pregnancy, Childbirth, and the Puerperium: 10.7%
    - Newborns and Other Neonates with Conditions Originating in the Perinatal Period: 9.5%
    - Diseases and Disorders of the Musculoskeletal System and Connective Tissue: 8.8%
    - Diseases and Disorders of the Digestive System: 8.5%



12. What is the percentage of apr_severity_of_illness_descriptions among patients?
    - Moderate: 38.39%
    - Minor: 33.17%
    - Major: 22.15%
    - Extreme: 6.29%

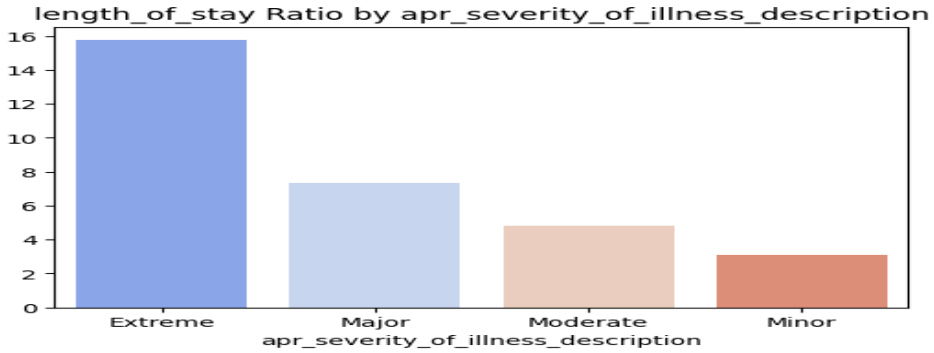


13. What is the percentage of payment_typology_1 among patients?
    - Medicare: 37.56%
    - Medicaid: 30.36%
    - Private Health Insurance: 14.22%
    - Blue Cross/Blue Shield: 11.75%
    - Self-Pay: 2.89%
    - Miscellaneous/Other: 1.24%
    - Managed Care, Unspecified: 0.98%
    - Federal/State/Local/VA: 0.53%
    - Unknown: 0.27%
    - Department of Corrections: 0.19%



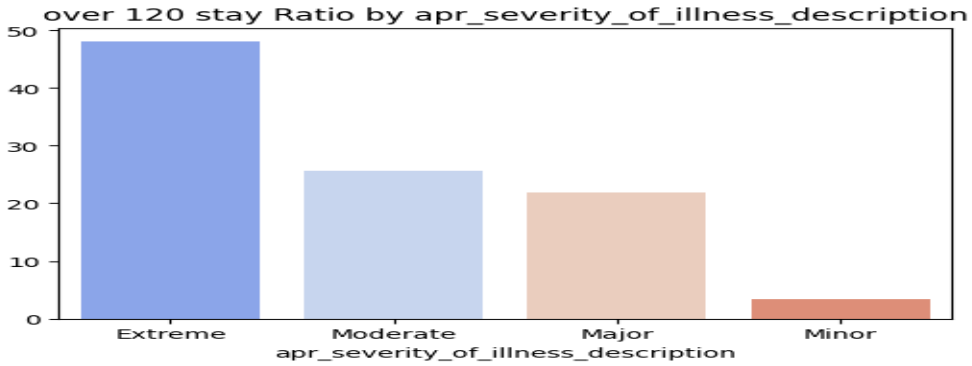
14. Calculate the average length_of_stay ratio for apr_severity_of_illness_description:
    - Extreme: 15.75

- Major: 7.33
- Moderate: 4.83
- Minor: 3.09



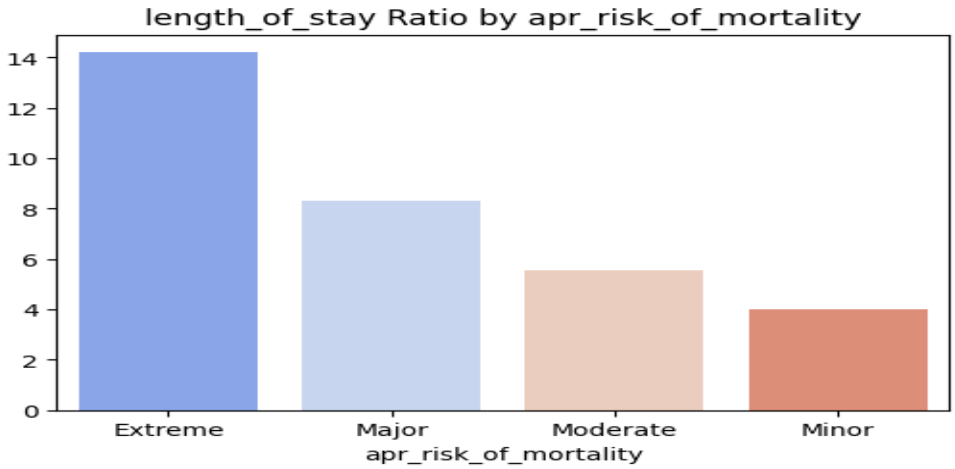
- From the above, it is clear that as the severity of illness increases, there is an increase in the length of stay days, which is logical.

15. What is the impact of severity of illness by patient stay for over 120 days?
    - Extreme 48.05
    - Moderate 25.71
    - Major 21.92
    - Minor 3.36



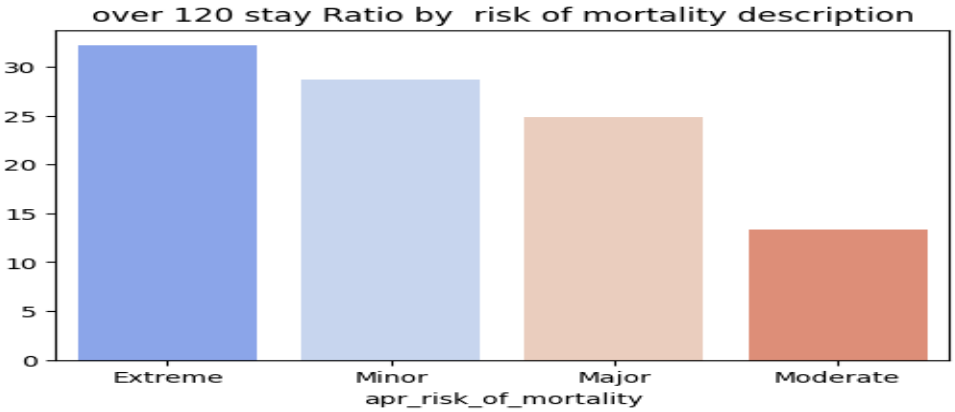
- From the above, we can observe the distribution, in percent, of long stays exceeding 120 days by severity of illness. The highest percentage, 48.05%, is for Extreme Severity, out of a total of 1857 patients with stays over 120 days.

16. Calculate the average length_of_stay ratio for apr_risk_of_mortality:
    - Extreme: 14.20
    - Major: 8.28
    - Moderate: 5.56
    - Minor: 4.02



- From the above, it is clear that as the risk of mortality increases, there is an increase in the length of stay days, which is logical.

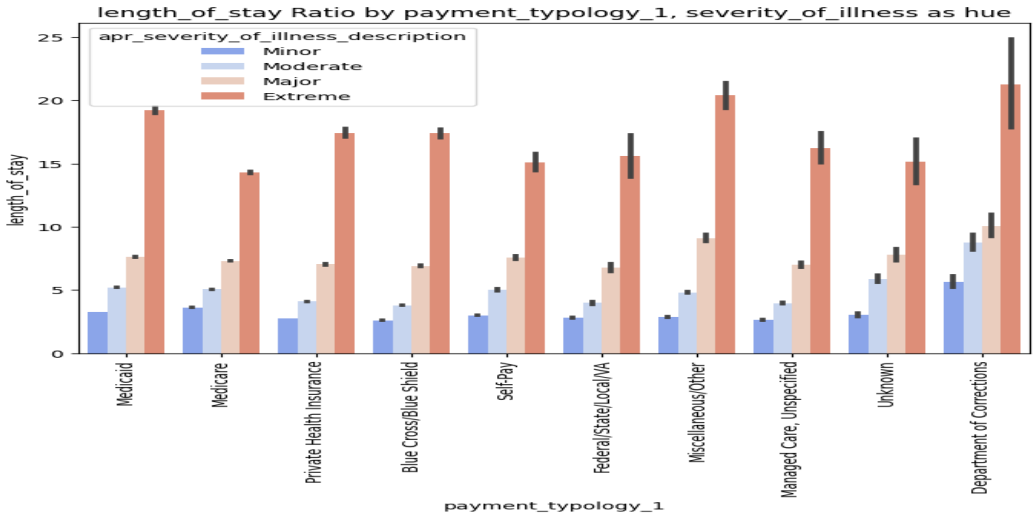
17. Calculate average over 120 stay Ratio for apr_risk_of_mortality?
    - Extreme 32.16
    - Minor 28.69
    - Major 24.85
    - Moderate 13.33



- From the above, we can observe the distribution, in percent, of long stays exceeding 120 days by risk of mortality. The highest percentage 32.16%, is for Extreme risk of mortality, out of a total of 1857 patients with stays over 120 days.

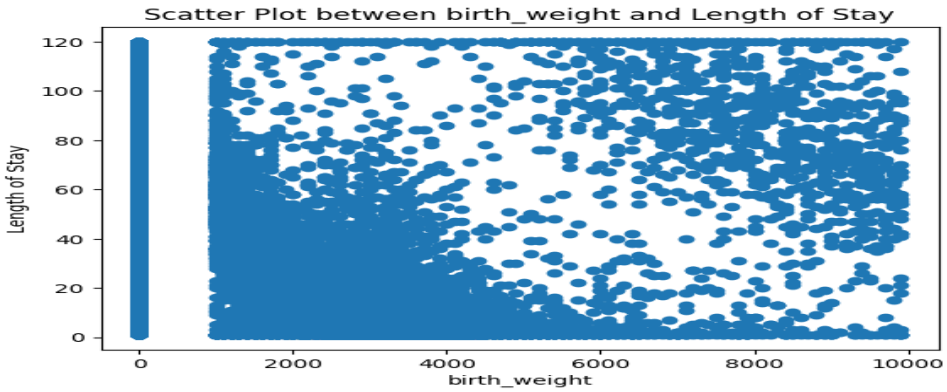
18. Calculate the average length_of_stay ratio for apr_medical_surgical_description:
    - Surgical: 5.99
    - Medical: 5.34
    - From the above, both medical and surgical descriptions have almost the same effect on the length of stay.

19. Calculate the average length_of_stay ratio for payment_typology_1:
    - Department of Corrections: 8.33
    - Medicare: 6.59
    - Unknown: 5.81
    - Miscellaneous/Other: 5.61
    - Medicaid: 5.29
    - Self-Pay: 4.98
    - Managed Care, Unspecified: 4.41
    - Federal/State/Local/VA: 4.35
    - Private Health Insurance: 4.31
    - Blue Cross/Blue Shield: 4.17



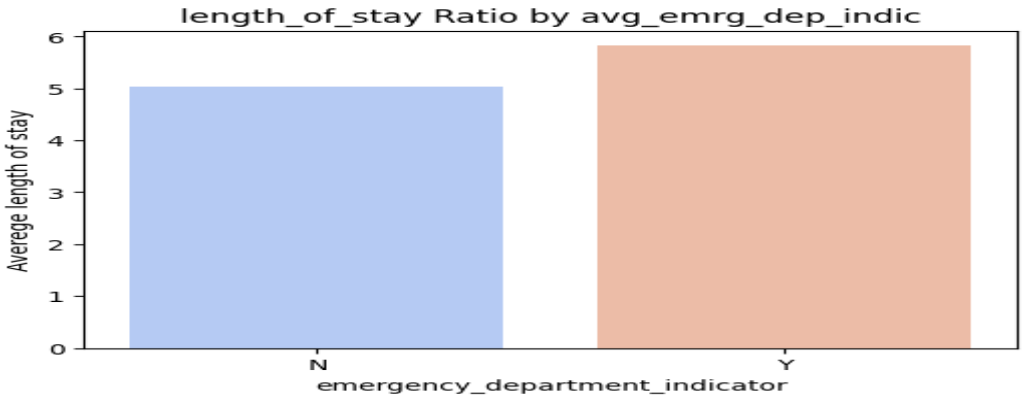
- Although the percentage of payment_typology_1 among patients shows that Medicare has the highest percentage, almost 38%, while Department of Corrections has the lowest percentage, at 0.19%.
- The above-average length_of_stay ratio for payment_typology_1 shows that the highest mean length of stay is for Department of Corrections, with an average of more than 8 days, approximately two days longer than Medicare.
- When investigating this contradiction further by looking at the length_of_stay ratio by payment_typology_1 and severity_of_illness as a hue, it becomes apparent that all severity_of_illness types are much more frequent in Department of Corrections than in other payment_typology_1 types. This may be one of the main reasons that could justify the contradiction, as severity_of_illness has a positive correlation with the length of stay.

20. What is the relationship between birth weight and length of stay?



- From the correlation heatmap, it appears that there is a negative correlation between birth weight and length of stay.
- Additionally, the scatter plot illustrates that as birth weight decreases, the data becomes more skewed and condensed to the left.

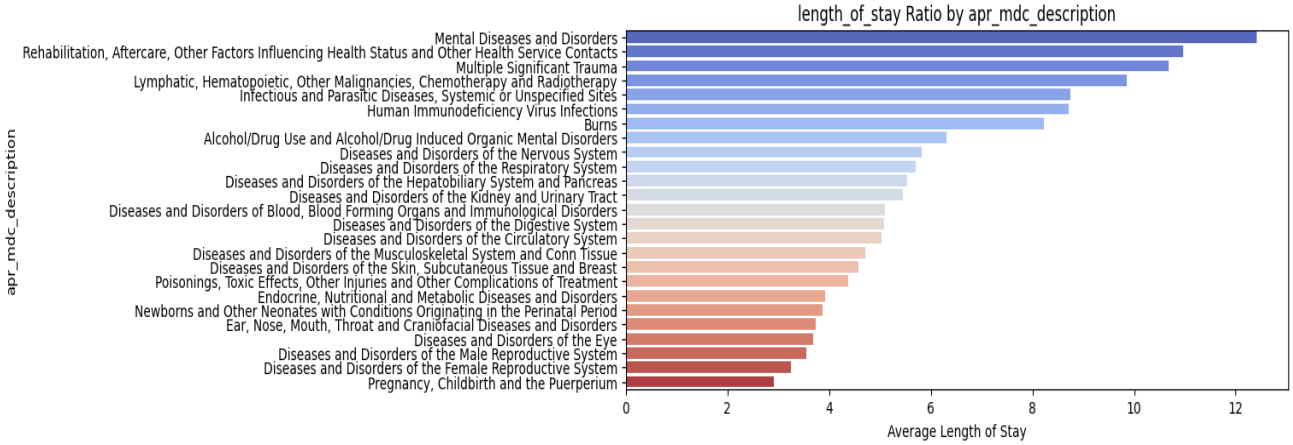
21. What is the effect of emergency_department_indicator on length_of_stay ratio?
    - N: 5.03
    - Y: 5.82



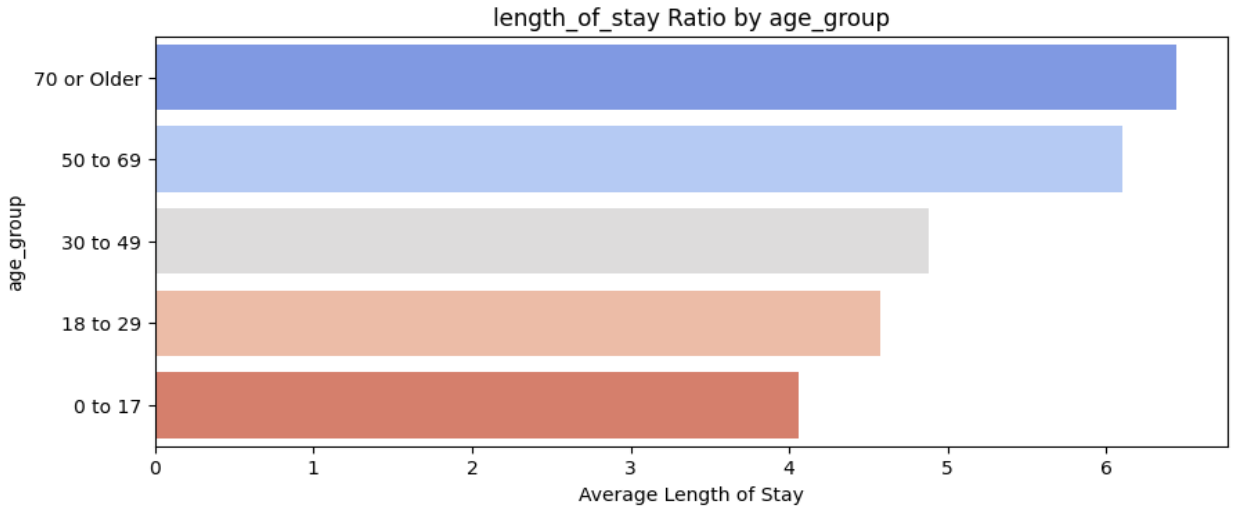
- From the above, the emergency_department_indicator of type Y (yes) has a higher average length of stay by almost 1 day more.

22. Calculate the average length_of_stay ratio for apr_mdc_description:
    - Mental Diseases and Disorders: 12.41
    - Rehabilitation, Aftercare, Other Factors Influencing Health Status and Other Health Service Contacts: 10.96
    - Multiple Significant Trauma: 10.68
    - Lymphatic, Hematopoietic, Other Malignancies, Chemotherapy and Radiotherapy: 9.86
    - Infectious and Parasitic Diseases, Systemic or Unspecified Sites: 8.75
    - Human Immunodeficiency Virus Infections: 8.72

- Burns: 8.22
- Alcohol/Drug Use and Alcohol/Drug-Induced Organic Mental Disorders: 6.31
- Diseases and Disorders of the Nervous System: 5.83
- Diseases and Disorders of the Respiratory System: 5.70
- Diseases and Disorders of the Hepatobiliary System and Pancreas: 5.54
- Diseases and Disorders of the Kidney and Urinary Tract: 5.45
- Diseases and Disorders of Blood, Blood-Forming Organs, and Immunological Disorders: 5.09
- Diseases and Disorders of the Digestive System: 5.08
- Diseases and Disorders of the Circulatory System: 5.03
- Diseases and Disorders of the Musculoskeletal System and Connective Tissue: 4.71
- Diseases and Disorders of the Skin, Subcutaneous Tissue, and Breast: 4.59
- Poisonings, Toxic Effects, Other Injuries, and Other Complications of Treatment: 4.38
- Endocrine, Nutritional, and Metabolic Diseases and Disorders: 3.92
- Newborns and Other Neonates with Conditions Originating in the Perinatal Period: 3.87
- Ear, Nose, Mouth, Throat, and Craniofacial Diseases and Disorders: 3.73
- Diseases and Disorders of the Eye: 3.69
- Diseases and Disorders of the Male Reproductive System: 3.55
- Diseases and Disorders of the Female Reproductive System: 3.25
- Pregnancy, Childbirth, and the Puerperium: 2.91

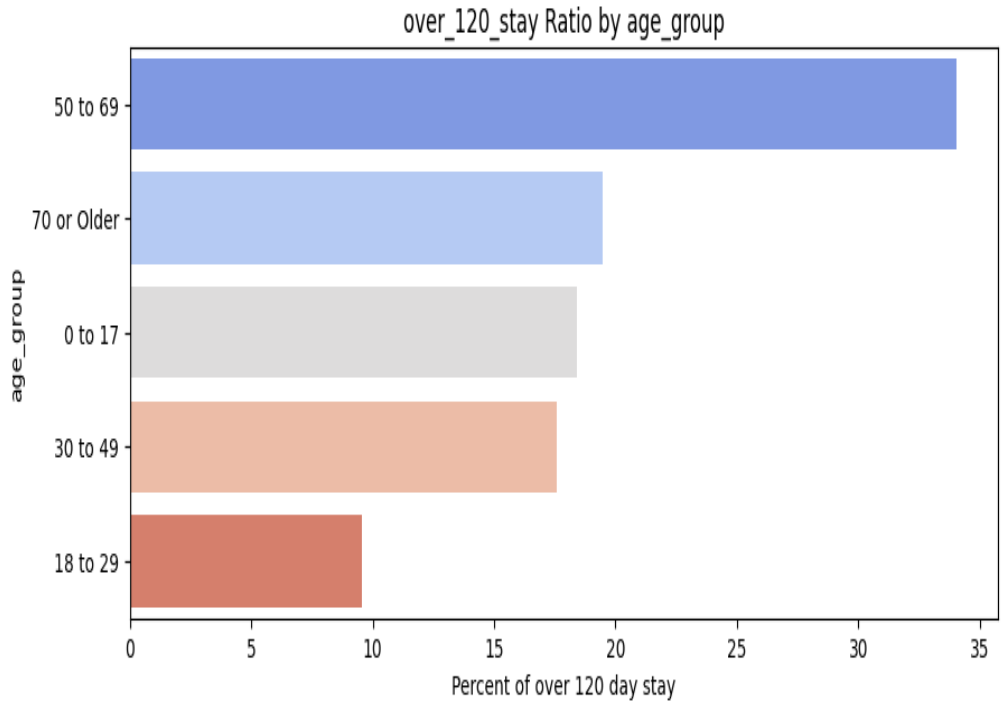


23. Calculate the average length_of_stay ratio for age_group:
    - 70 or Older: 6.44
    - 50 to 69: 6.10
    - 30 to 49: 4.88
    - 18 to 29: 4.58
    - 0 to 17: 4.06



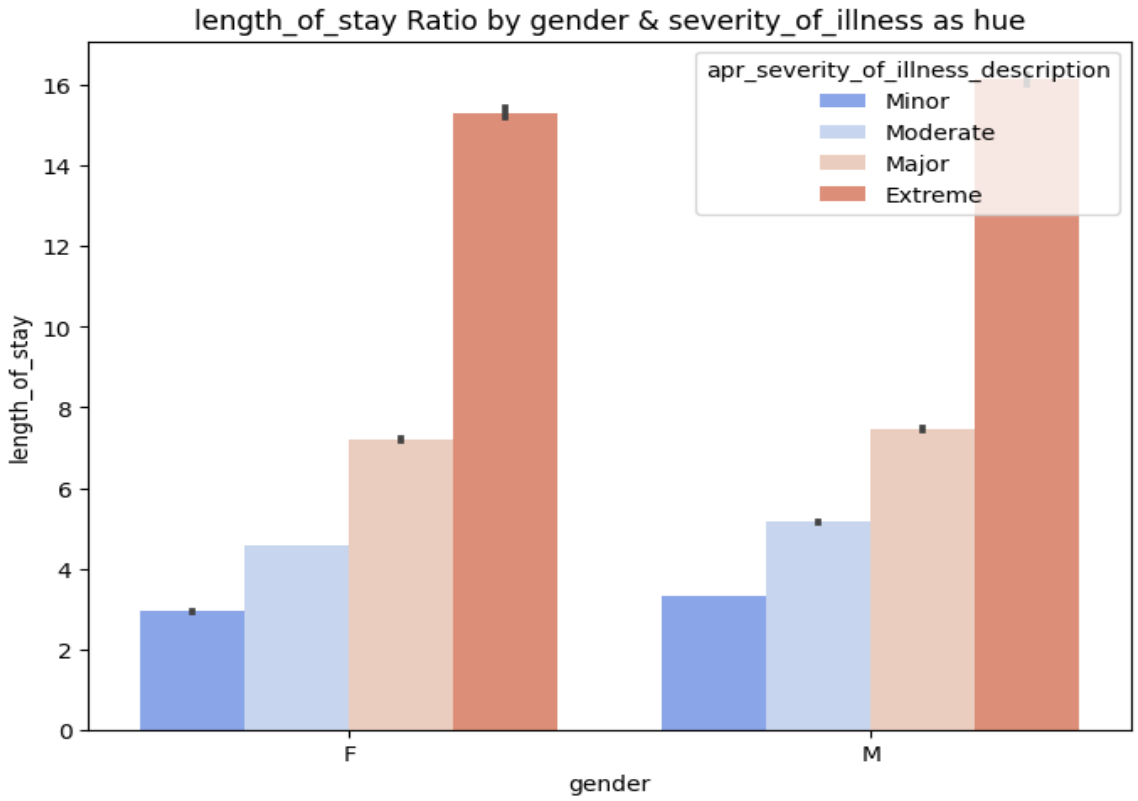
- From the above, we find that the younger age group tends to have a lower average length of stay than the older age group.

24. Checking the effect of age group by over_120_stay feature:
    - age_group
    - 50 to 69 34.03
    - 70 or Older 19.47
    - 0 to 17 18.40
    - 30 to 49 17.60
    - 18 to 29 9.55



- From the above, we can observe the distribution, in percent, of long stays exceeding 120 days by age group. The highest percentage, 34.03%, is for individuals aged between 50 and 69 years old, out of a total of 1857 patients with stays over 120 days.

25. Calculate the average length_of_stay ratio for gender:
    - Male: 5.97
    - Female: 5.12



- Although females represent 55% of the total patients, they show a lower average length of stay.
- When investigating this contradiction further by looking at the length_of_stay ratio by gender and severity_of_illness as a hue, it becomes apparent that all severity_of_illness types are higher in frequency in males than in females. This may be one of the main reasons that could justify the contradiction, as severity_of_illness has a positive correlation with the length of stay.

In conclusion of Analysis step, the analysis of patient data in New York State for the year 2015 revealed significant insights. Severity of illness and risk of mortality correlate with longer hospital stays. Payment type, hospital location, and demographics impact length of stay. Additionally, a negative correlation exists between birth weight and length of stay. These findings provide valuable insights for healthcare planning and resource allocation.

Step3: Data Preprocessing:

Preprocessing Outcomes Summary:

The main objective of preprocessing and outcomes as follows:

1. Checking the dataset for irrelevant columns that do not serve our prediction model of length of stay and dealing with them.

- Not all the features serve our aim, as we attempt to create a general model that can predict patient length of stay regardless of the city or state. Additionally, 'total cost' and 'charges' exceeding 120 days of stay do not make sense to keep.

The following columns were removed:

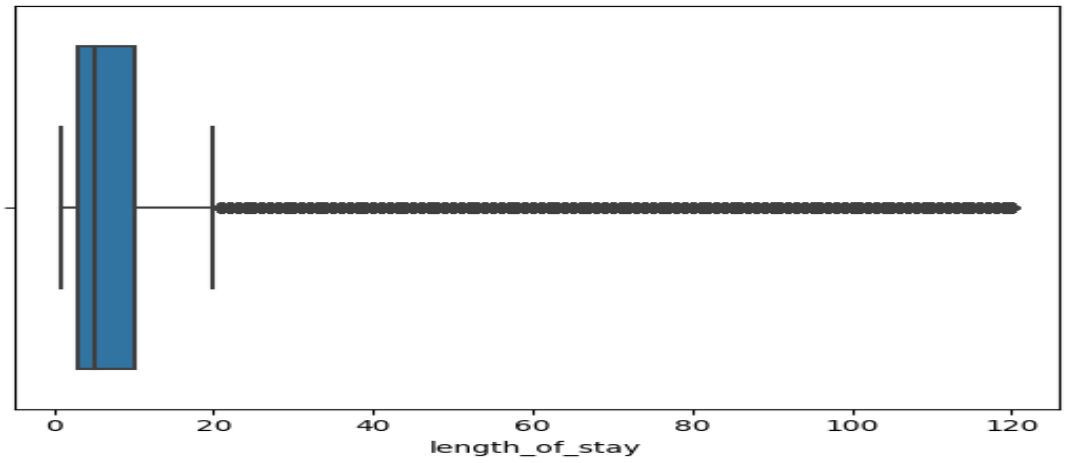
['health_service_area','hospital_county','operating_certificate_number','facility_id','facility_name','race','ethnicity', 'zip_code__3_digits', 'patient_disposition', 'ccs_diagnosis_code', 'ccs_procedure_code', 'apr_drg_code',' apr_mdc_code', 'apr_severity_of_illness_code', 'payment_typology_1', 'birth_weight','total_charges', 'total_costs','over_120_stay'].

- After dropping the above irrelevant features 1,264,916 duplicates appears in the data, we drop all of them.

2. Selecting & dropping not adult patients as their treatment may vary a lot from adult.

- we drop 342237 instants of 0 to 17 patients and we drop all of them to maintain the data for adult patients.

3. handling outliers:



There were outliers above 20 days we deal with them by change their values to the end of the Wisker which is almost 20 days by this way we are going to keep those patient data and point that prediction of 20 days it means 20 days and above.

4. Handling Categorical Variables:

1) For ordinal categorical features:

- By creating dictionary with ordinal number and mapping the dictionary with the ordinal categorical feature.

2) Encoding nominal feature:

* For Features with high cardinality we use Target Encoding (Mean Encoding) for nominal categorical features then dropping it features like:

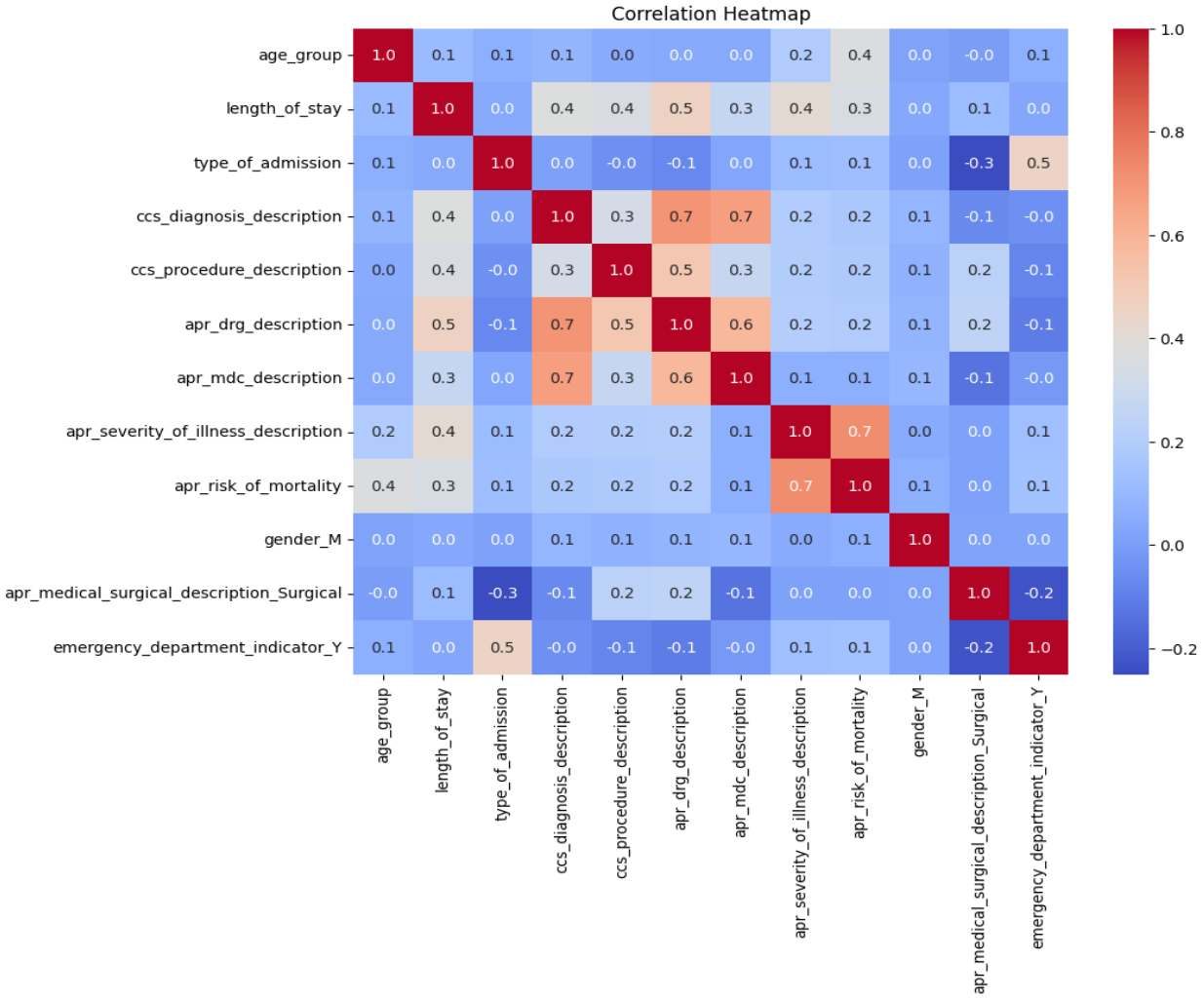
- (ccs_diagnosis_description: 257 unique values, ccs_procedure_description: 232, apr_drg_description: 285, apr_mdc_description: 24, type_of_admission: 6 unique values)

* For features with low cardinality we use pandas get dummies to encode it features like:

- (gender: 2 unique, apr_medical_surgical_description: 2, emergency_department_indicator: 2 unique values)

5. Correlation heatmap: Checking correlation matrices for the target length of stay after encoding the features. The results were as follows:

- length_of_stay	1.00
- apr_drg_description	0.46
- apr_severity_of_illness_description	0.41
- ccs_diagnosis_description	0.37
- ccs_procedure_description	0.35
- apr_risk_of_mortality	0.35
- apr_mdc_description	0.27
- apr_medical_surgical_description_Surgical	0.11
- age_group	0.11
- type_of_admission	0.04
- gender_M	0.04
- emergency_department_indicator_Y	0.03



- After checking the correlation heat map there are no highly correlated features above 0.8 correlation so, we are going to keep all the features.

* Saving a dictionary for categorical features that will be needed in the model deployment step.

6. After checking for duplication after all of the above process we found:

1) After step 1 dropping the irrelevant features, (1,264,916) duplicates appears in the data we drop them.

2) After step 2 we drop (342,237) instants of 0 to 17 patients to maintain the data sets for adult patients.

3) checking duplicates after handling outliers & encoding (26,320) duplicated value. we drop all of it and rest the index.

7. Saving the preprocessed, encoded, and scaled CSV file for the next modeling step.

Step4: Modeling:

Modeling Main Objectives:

1. Maintain the huge size of the dataset, which will be challenging when applying different modeling approaches.
2. Apply three different models and compare their results using two evaluation metrics:
    - Use model accuracy scores ranging from 0 to 1, where a higher score indicates a better fit.
    - Utilize Mean Absolute Error (MAE), a measure of the average absolute difference between predicted and actual values.
    - Compare the model accuracy score to MAE scores for different models. The model with the lowest MAE and the highest accuracy score will be selected.
    - Check the importance of the features for the model.
3. Tune the algorithm selected as the best model with the highest evaluations from the above steps using cross-validation to select the best parameters for the model.

Maintaining the Huge Size of the Dataset:

Despite the data size decreasing to 722,453 entries and 12 features after cleaning and preprocessing, dealing with it on a limited processor laptop remains challenging. To overcome this, a small sample was taken to run the models more efficiently while maintaining the same characteristics of the data using the stratify method.
During the trials:

1. Standard Scaler was used to standardize the independent variables, but robust scaler was also checked with no changes in accuracy.

2. Different regressors and classification models were tested, but all showed poor accuracy results.

3. Feature importance was checked using the random forest model, and while dropping less important features was attempted, no enhancements appeared in model accuracy.

4. Polynomial Features were tried with different degrees, and no improvements were observed.

5. Principal Component Analysis (PCA) was applied after polynomial features with varying component numbers, but nothing changed.

6. Considering the models' poor performance, an attempt was made to treat the problem as a classification model with the target class between 1 to 20 days as labels. The random forest classifier showed good accuracy (96%) in the train set for recall and precision scores but performed poorly (14%) on the test set.

7. Cross-validation and Grid Search CV were applied for hyperparameter tuning, but no significant changes occurred.

8. Due to highly imbalanced data, different techniques of under-sampling and over-sampling were tried: A. Using random under-sampler showed no noticeable change. B. Using SMOTE as an over-sampler improved the results. C. Using random over-sampler significantly improved the results.

9. After addressing the imbalance problem, regression models were retested, showing high accuracy results. Models were also tested with length of stays from 1 to 20 days and total length of stay from 1 to 120 days.

10. Then final model was deployed using Streamlet.

Model & Project Summary:

- Although the model demonstrated an impressive training and testing accuracy of 99%, the results were based on a minimal 0.01% sample size of the original data, carefully selected using stratified sampling. While this model shows promise, it cannot yet be considered ready for generalization. To achieve that level of trust and applicability, further investment in terms of time, computational resources, and extensive studies is required.

- Throughout this project, numerous techniques were explored in several different notebooks of preprocessing & modeling, encompassing feature selection, outlier handling, diverse encoding methods for various classification and regression models, and cross-validation techniques. These endeavors represent a comprehensive effort to develop a robust predictive model for inpatient length of stay.

- The journey has been informative, revealing both the potential of machine learning in healthcare and the challenges associated with working on real-world, large-scale datasets. Continued research, collaboration, and refinements are essential for building a truly reliable and widely applicable solution.