**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 have '120 +' value repeated 1857 time out of above 3 million record I treated all enters '120 +' as 120. also, dtype changed to numerical.
4. discharge\_year feature has only 2015 value. it will not be useful so drop it.
5. attending\_provider\_license\_number feature it is not useful and we drop it.
6. abortion\_edit\_indicator feature has only 'N' value. it will not be useful we drop it.
7. total\_charges feature has Dollar signs. Sign removed and replaced dtype to numerical.
8. total\_costs feature has Dollar signs. Sign removed and replaced dtype to numerical.

* 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:**

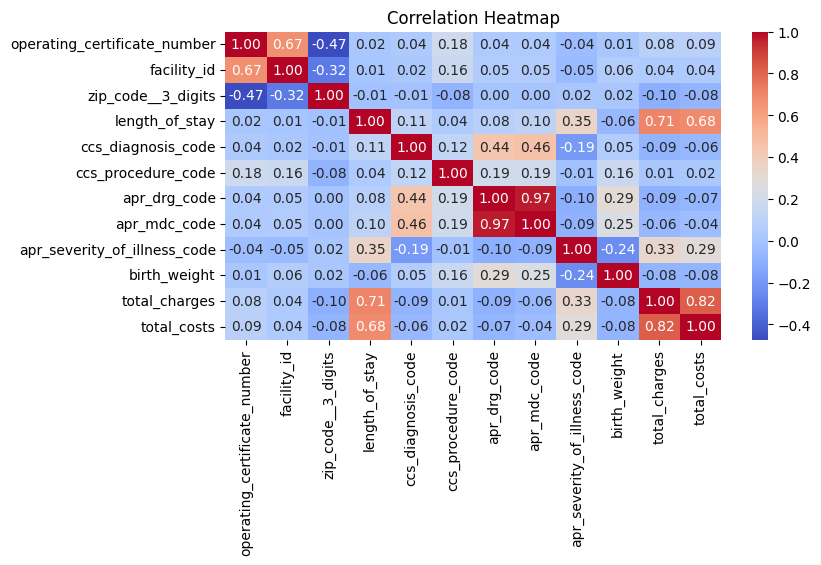
1. What is the patient distribution of most important features of New York State during the 2015-year dataset?
2. Calculate average Length of Stay Ratio for top or all if applicable features (APR Severity of Illness Code, APR Risk of Mortality, APR Medical Surgical Description, Payment Typology 1, Emergency 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.

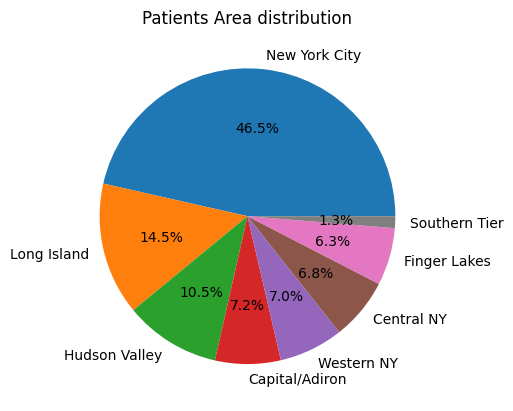
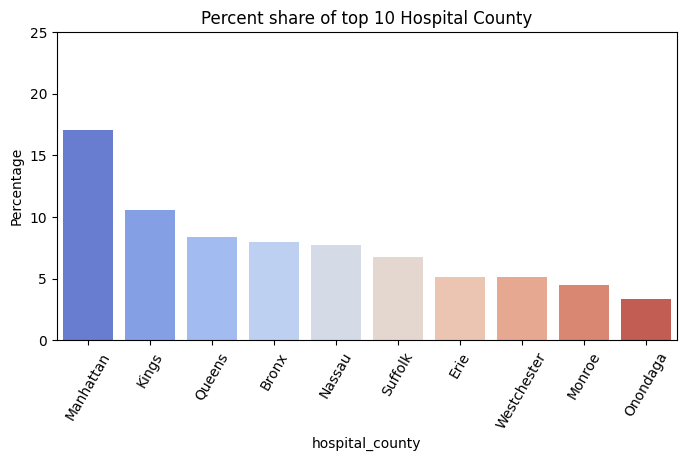
* **First Statistical Description for Numeric Features for the State of New York 2015 Inpatient Discharges:** We are going to focus on important numerical features:
  1. **Length\_of\_Stay Feature:** Ranges from 1 to 120 days. The mean length of stay is almost 6 days for the entire population, and the 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.

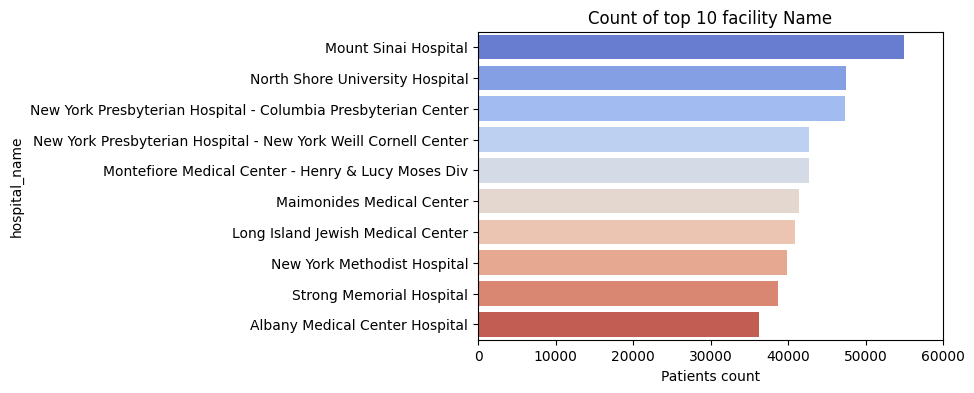
**2. The Correlation & Correlation heatmap outcomes:** for important numerical feature.

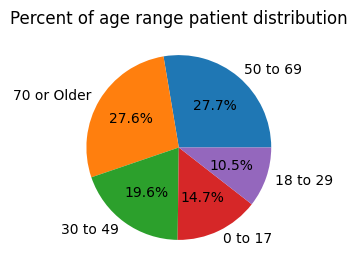
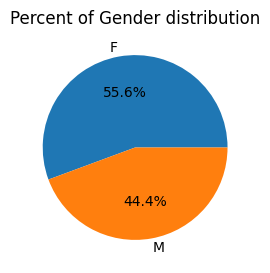
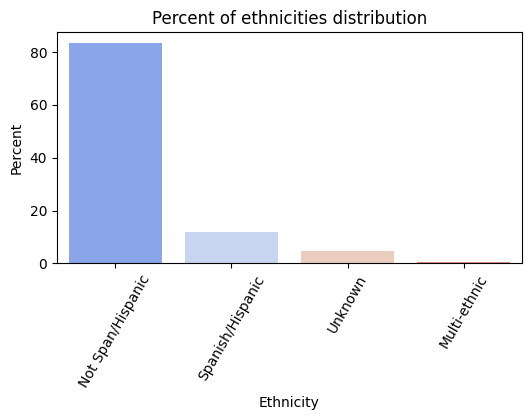


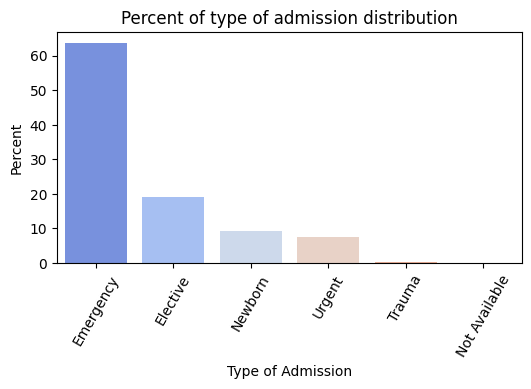
Although the correlation matrix may not provide meaningful insights for all columns, it is evident that the APR Severity of Illness Code shows a strong positive correlation with the length of stay, as do total charges and total costs. Additionally, CCS Diagnosis code appears to have a slight positive 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:**

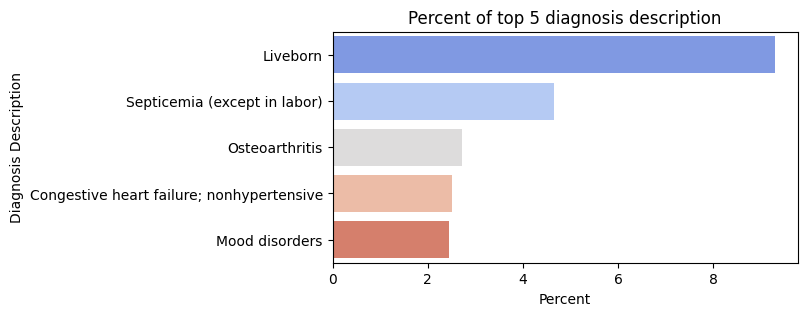
1. What is the distribution percentage of patients among cities of New York State during the year 2015?
   * New York City: 46.46%
   * Long Island: 14.50%
   * Hudson Valley: 10.51%
   * Capital/Adirondack: 7.17%
   * Western NY: 7.02%
   * Central NY: 6.77%
   * Finger Lakes: 6.28%
   * Southern Tier: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



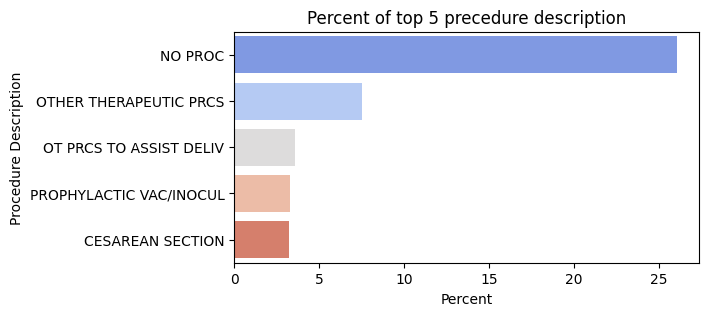
1. What is the distribution percentage of age groups among patients during the year 2015?
   * 50 to 69: 27.65%
   * 70 or Older: 27.58%
   * 30 to 49: 19.56%
   * 0 to 17: 14.68%
   * 18 to 29: 10.53%
2. What is the percentage of gender distribution among patients?
   * Female: 55.64%
   * Male: 44.36%
3. 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%
4. 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%



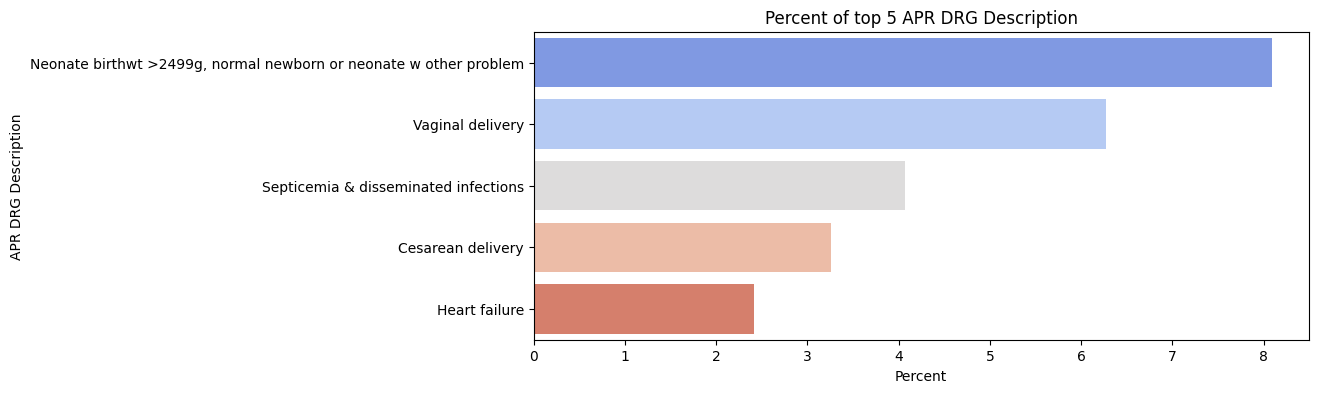
1. What are the top 5 diagnosis descriptions among patients, and what percentage does each one represent?
   * Liveborn: 9.31%
   * Septicemia (except in labor): 4.66%
   * Osteoarthritis: 2.72%
   * Congestive heart failure; nonhypertensive: 2.52%
   * Mood disorders: 2.44%



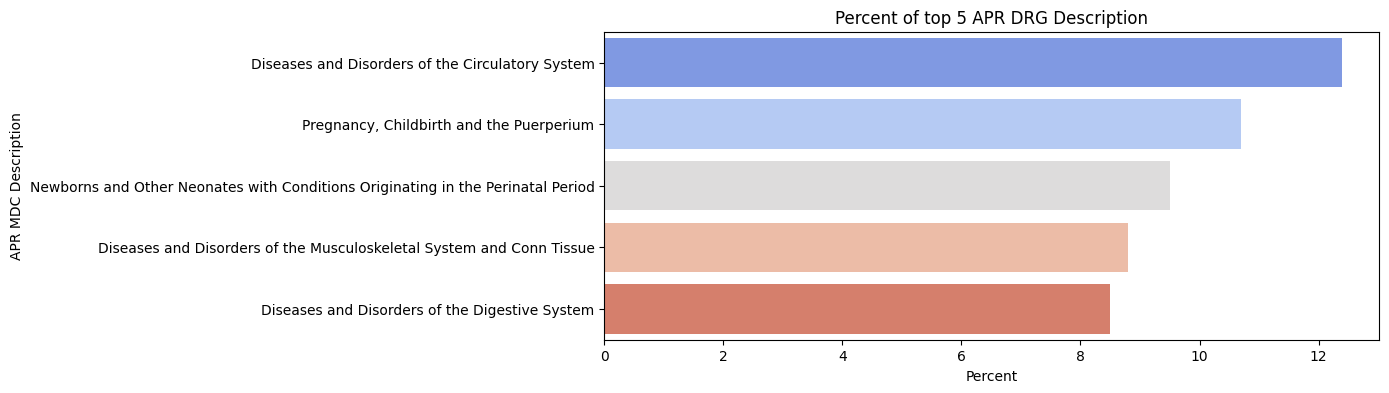
1. What are the top 5 procedure descriptions among patients, and what percentage does each one represent?
   * NO PROC: 26.05%
   * OTHER THERAPEUTIC PRCS: 7.51%
   * OT PRCS TO ASSIST DELIV: 3.56%
   * PROPHYLACTIC VAC/INOCUL: 3.28%
   * CESAREAN SECTION: 3.20%



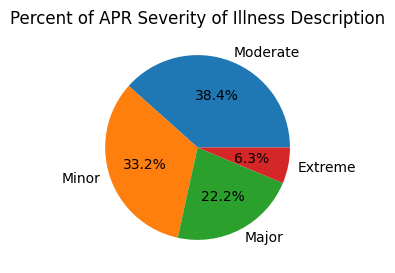
1. What are the top 5 apr\_drg\_descriptions among patients, and what percentage does each one represents?
   * Neonate birth weight >2499g, normal newborn or neonate w other problem: 8.09%
   * Vaginal delivery: 6.27%
   * Septicemia & disseminated infections: 4.07%
   * Cesarean delivery: 3.25%
  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



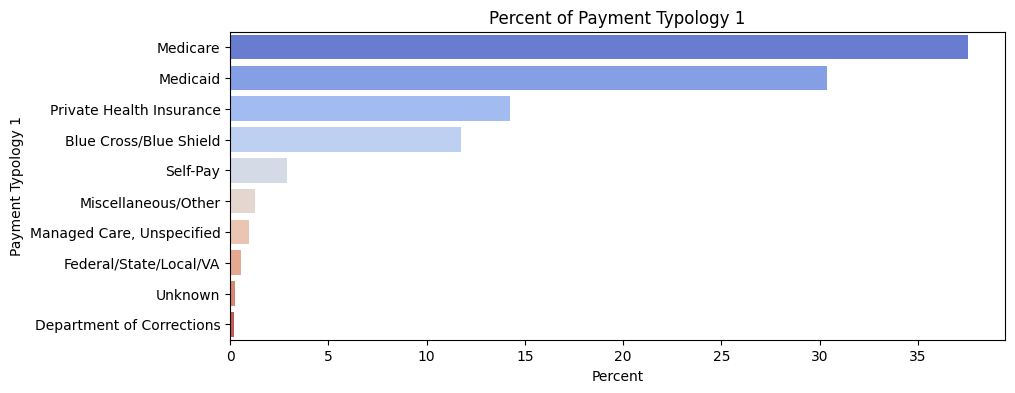
1. What are the top 5 apr\_mdc\_descriptions among patients, and what percentage does each one represent?
   * Diseases and Disorders of the Circulatory System: 12.4%
   * Pregnancy, Childbirth, and the Puerperium: 10.7%
   * Newborns and Other Neonates with Conditions Originating in the Perinatal Period: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



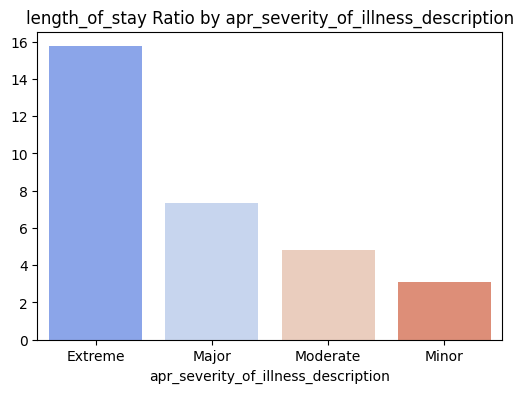
1. What is the percentage of apr\_severity\_of\_illness\_descriptions among patients?
   * Moderate: 38.39%
   * Minor: 33.17%
   * Major: 22.15%
   * Extreme: 6.29%



1. 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%

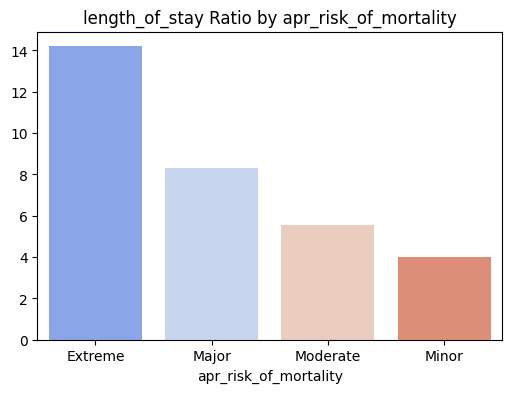


1. Calculate the average length\_of\_stay ratio for apr\_severity\_of\_illness\_description:
   * Extreme: 15.75
   * Major: 7.33
   * Moderate: 4.83
   * Minor: 3.09



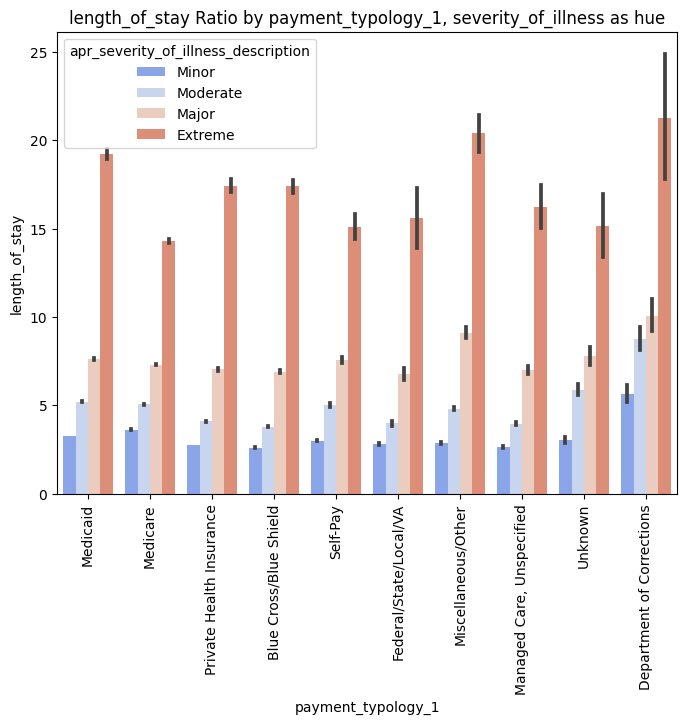
* + From the above, it is clear that as the severity of illness increases, there is an increase in the length of stay days, which is logical.

1. Calculate the average length\_of\_stay ratio for apr\_risk\_of\_mortality:
   * Extreme: 14.20
   * Major: 8.28
   * Moderate: 5.56
   * Minor: 4.02



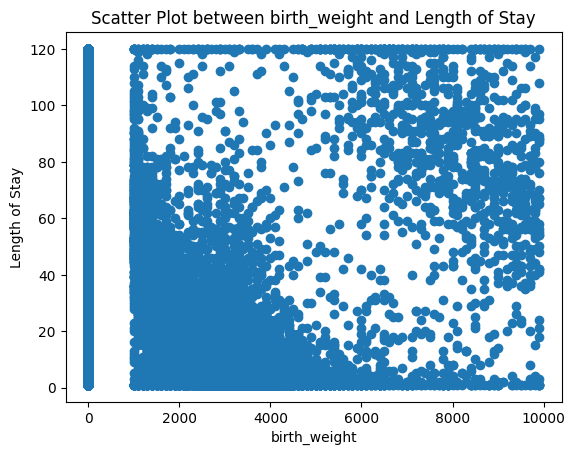
* + From the above, it is clear that as the risk of mortality increases, there is an increase in the length of stay days, which is logical.

1. 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.
2. 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

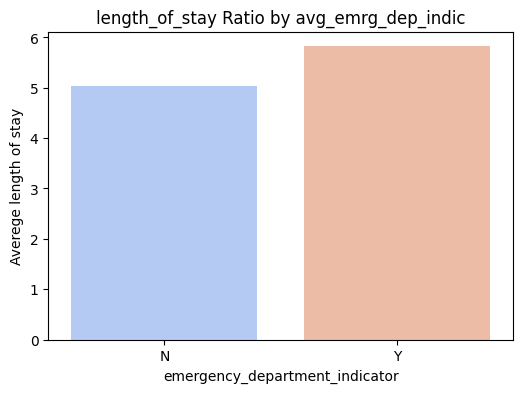


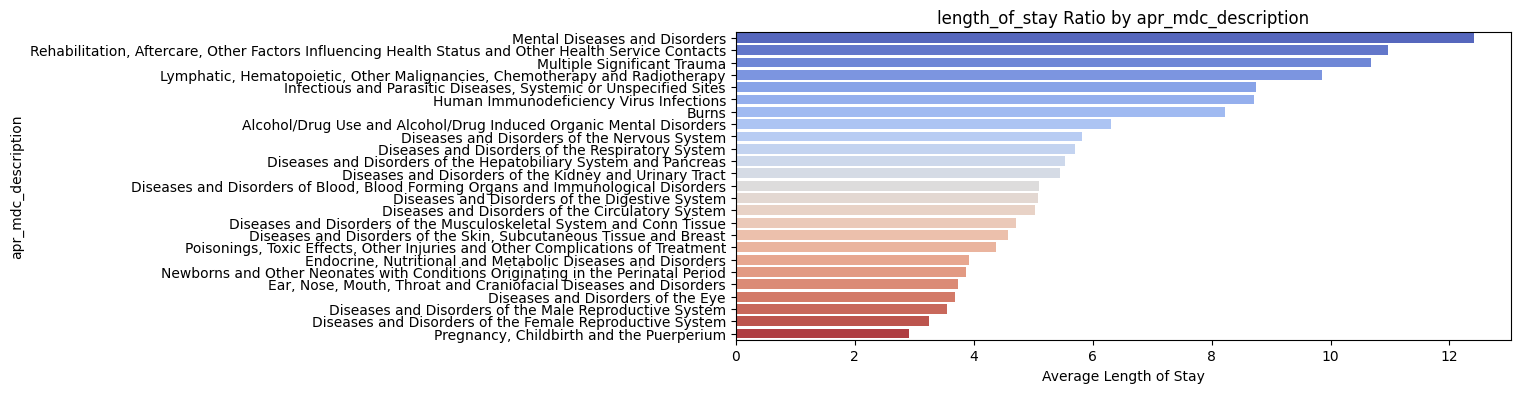
* + Although the percentage of payment\_typology\_1 among patients shows that Medicare has the highest percentage, almost 38%, while Department of Corrections has the lowest percentage, at 0.19%.
  + The above-average length\_of\_stay ratio for payment\_typology\_1 shows that the highest mean length 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.

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

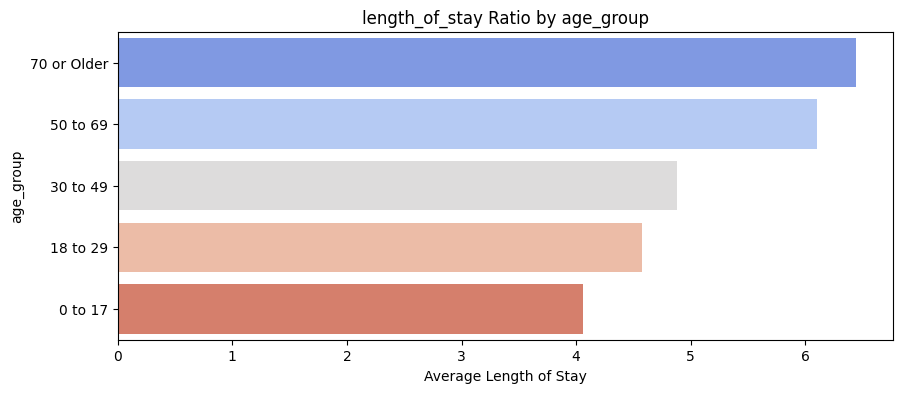


* + From the correlation heatmap, it appears that there is a negative correlation between birth weight and length of stay.
  + Additionally, the scatter plot illustrates that as birth weight decreases, the data becomes more skewed and condensed to the left.

1. 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.
2. 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

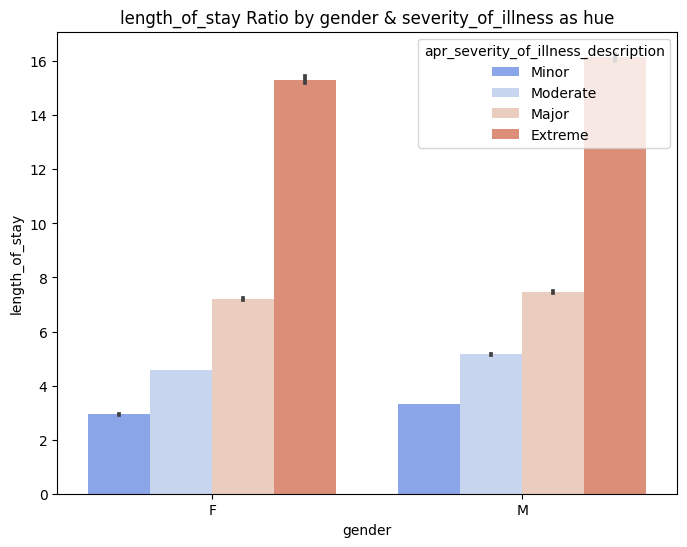


1. Calculate the average length\_of\_stay ratio for age\_group:
   * 70 or Older: 6.44
   * 50 to 69: 6.10
   * 30 to 49: 4.88
   * 18 to 29: 4.58
   * 0 to 17: 4.06



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

1. 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.