**Analyzing Metastatic Breast Cancer By Using Big Data**

Group 21: 519544 Yongfan Zhao, 530667 Jiarui Ma, 530365 Tong Hang

**Executive summary**

This report explores the relationships between medications, adverse reactions, and outcomes in patients with metastatic breast cancer. Using a large dataset of 72,403 unique patients, we analyzed treatment patterns, symptoms, and outcomes to address critical challenges in managing this severe disease.

Our findings reveal that **Ibrance** and **Kisqali** are among the most frequently prescribed drugs, with common adverse reactions such as fatigue and neutropenia. Importantly, we found no significant differences in patient outcomes when comparing pre-and post-COVID-19 cases, highlighting the resilience of treatment protocols during the pandemic.

Methodologically, we applied logistic regression to predict patient outcomes, achieving an accuracy of 62.53%. However, the complexity of the data—characterized by high-cardinality categorical variables—limited the model’s performance. Attempts to use random forest classifiers were hindered by computational constraints, emphasizing the need for advanced tools and techniques in future research.

This study underscores the challenges of handling large, complex datasets in healthcare analytics. Key recommendations include developing better methods for processing string and categorical data, as well as investing in computational resources to support sophisticated modeling approaches. These steps will be vital for improving predictive accuracy and optimizing patient care in metastatic breast cancer treatment.

**Description of the data**

Our dataset focuses on patients with metastatic breast cancer. It includes each patient’s age, weight, gender, COVID-19 status, current symptoms, and medications used. Each symptom is documented on a separate row in the PT column using application-specific terminology when recording a patient's symptoms.

The dataset consists of 11 columns and 1,334,114 rows (Appendix 1). However, this does not mean there are 1,334,114 individual patients. Since symptoms are recorded on separate rows, the dataset contains data for 72,403 unique patients (Appendix 2), of whom 71,686 are female and 717 are male, yielding a gender ratio of approximately 100:1 (Appendix 3).

During our analysis of missing values, we observed significant missing data in the age, weight (WT), and outcome code (OUTC\_COD) columns. In contrast, the PT column had only a small proportion of missing values (Appendix 4). We filled in the missing values in the age and weight columns to address these issues using the average values calculated separately for male and female patients(Appendix 5). For the OUTC\_COD column, we filled in missing values with "OT." However, we could not find a reasonable method to impute missing values in the PT column. As a result, we chose to drop 1,889 rows with missing PT data. We performed a verification analysis after filling in the missing values (Appendix 6).

**Problem Statement**

How do different drugs used in metastatic breast cancer treatment relate to adverse reactions and their patterns?

**Our primary aim is to:**

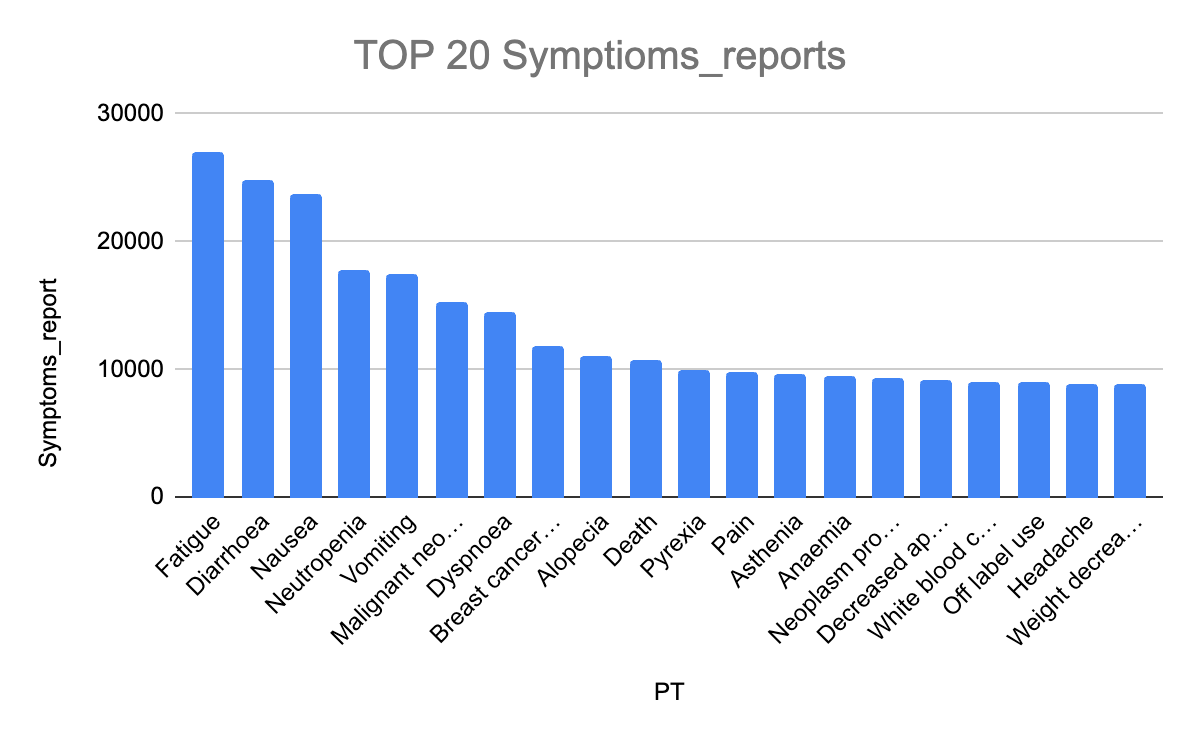
1. Identify the most frequently prescribed drugs for metastatic breast cancer patients.
2. Analyze the symptoms associated with these drugs and their frequency.
3. Investigate patterns of adverse reactions based on demographics (e.g., age, gender) and COVID-19.

**Research Questions:**

1. Which drugs are most frequently used in metastatic breast cancer treatment(Appendix 8)?

| **DRUGNAME** | **Drug\_used** | **for\_F** | **for\_M** |
| --- | --- | --- | --- |
| KISQALI | 125587 | KISQALI | DOCETAXEL |
| IBRANCE | 115596 | IBRANCE | PACLITAXEL |
| ZOMETA | 77080 | ZOMETA | PERTUZUMAB |
| LETROZOLE. | 53952 | LETROZOLE. | TRASTUZUMAB |
| LETROZOLE | 40087 | LETROZOLE | ZOMETA |
| TRASTUZUMAB | 37864 | HERCEPTIN | AREDIA |
| HERCEPTIN | 35116 | AFINITOR | XELODA |
| AFINITOR | 34507 | TRASTUZUMAB | IBRANCE |
| PALBOCICLIB | 31400 | PALBOCICLIB | KADCYLA |
| DOCETAXEL | 30179 | TRASTUZUMAB. | TRASTUZUMAB. |

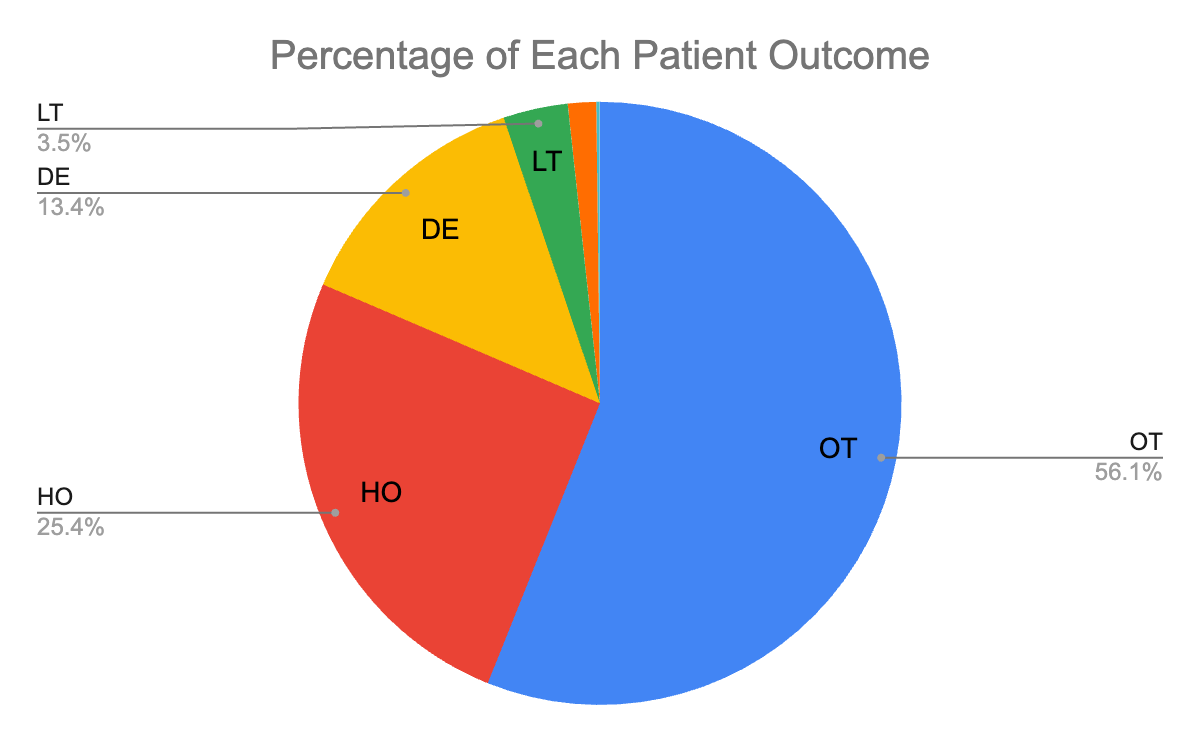
1. What are the most common symptoms or adverse reactions reported by metastatic breast cancer patients(Appendix 9)?



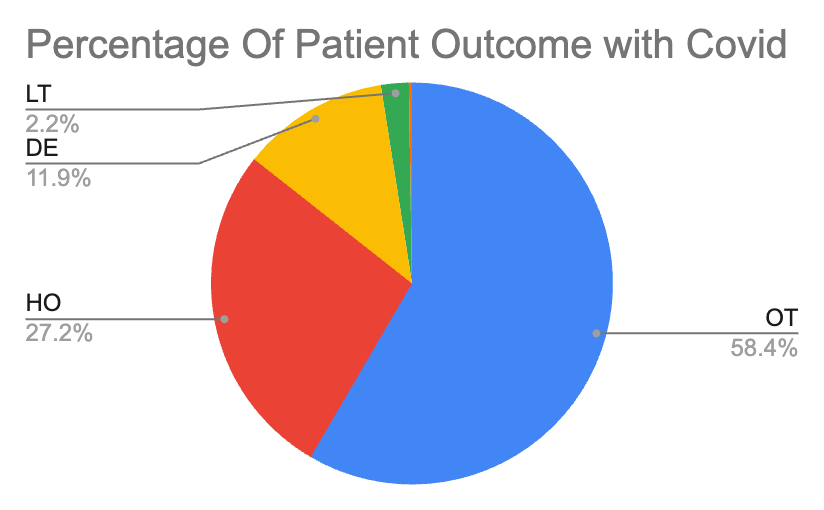
1. Are certain drugs associated with specific symptoms(Appendix 10)?

| **PT** | **DRUGNAME** | **reports** |
| --- | --- | --- |
| Fatigue | IBRANCE | 4809 |
| White blood cell count decreased | IBRANCE | 4038 |
| Neutropenia | KISQALI | 3665 |
| Neoplasm progression | IBRANCE | 3585 |
| Nausea | KISQALI | 3247 |
| Fatigue | KISQALI | 2715 |
| Leukopenia | KISQALI | 2631 |
| Nausea | IBRANCE | 2271 |
| Alopecia | IBRANCE | 2247 |
| General physical health deterioration | KISQALI | 2147 |

1. What is the mortality rate among metastatic breast cancer patients in the dataset? Are there any patterns linking specific adverse reactions or drug treatments to higher mortality rates(Appendix 11)?



1. Are there differences in reported patient Outcomes between pre- and post-COVID cases? Covid did not affect patient outcomes.



**Description of the disease**

What is the disease in focus?

Metastatic breast cancer is the most advanced stage of breast cancer, occurring when cancer cells spread beyond the breast to distant parts of the body, such as the bones, liver, lungs, or brain. It is also known as Stage IV breast cancer and is considered incurable. Still, advancements in treatment have made it manageable in many cases, with patients often living several years post-diagnosis.

What are the symptoms?

Back, bone, or joint pain, difficulty urinating, numbness or weakness anywhere in your body,

dry cough, shortness of breath, chest pain, loss of appetite, abdominal bloating, pain or tenderness, constant nausea, vomiting, or weight loss.

How many patients are in the US/world?

In the US:

Approximately 168,000 patients in the United States are living with metastatic breast cancer, including those diagnosed at this stage and those whose cancer has progressed from earlier stages.

Worldwide:

Breast cancer is the most commonly diagnosed cancer globally, with approximately 2.3 million cases annually. Among these, about 6–10% are metastatic at initial diagnosis, with many more progressing to metastatic disease after initial treatment.

Is it serious?

Yes, metastatic breast cancer is severe as it is life-threatening and requires lifelong treatment to manage symptoms and slow disease progression. The median survival for metastatic breast cancer has improved with advancements in treatment, but it remains approximately 2–5 years, varying based on factors like age, treatment, and the extent of metastasis.

Is there any specific part of the body involved?

Metastatic breast cancer can spread to several parts of the body, with the most common sites including:

Bones: The most frequent site of metastasis, often leading to pain and fractures.

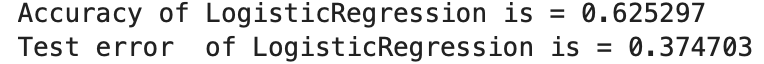
Lungs: Causes respiratory issues like shortness of breath and chest pain.

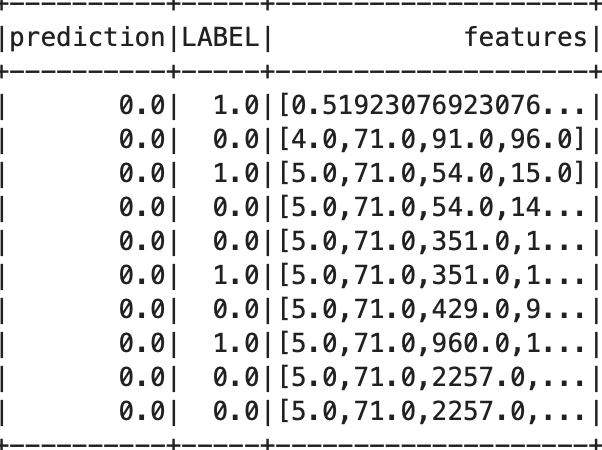
Liver: This may lead to jaundice, swelling, and abdominal discomfort.

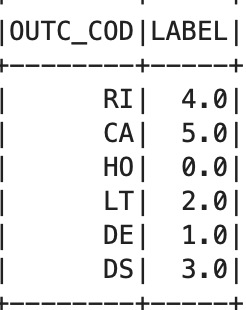
Brain: Less common but more severe, leading to headaches, seizures, and neurological deficits.

**Method & Results(Appendix 12)**

We attempted to use logistic regression to predict the patients' outcomes, selecting **age**, **weight**, **PT**, and **drug name** as features. To improve the dataset, we dropped rows where OUTC\_COD was labeled as "OT," as it lacks a clear or specific meaning. However, since **PT**, **drug name**, and patient outcomes are represented as strings, we first needed to convert these strings into numerical indices. This extensive conversion process introduced significant complexity, which likely reduced the accuracy of our final model.

We also experimented with a random forest classifier, but the diversity and complexity of the data made it computationally infeasible to run effectively. Below are the results of our logistic regression model, which achieved an accuracy of 62.53%, which suggests that the model captured some patterns in the data but faced challenges in handling the complexity of the relationship among the features. The test error of 37.47% indicates that over one-third of the predictions were incorrect. This performance could be improved with better feature representation or alternative models. The difficulty in accurately predicting outcomes can be due to high dimensionality, features like PT and DRUGNAME are categorical and have high cardinality, which complicates pattern recognition. And some outcome categories might have fewer instances, making it harder for model to learn these patterns effectively. If we implement more effective methods for handling string features, the model's accuracy could improve further.



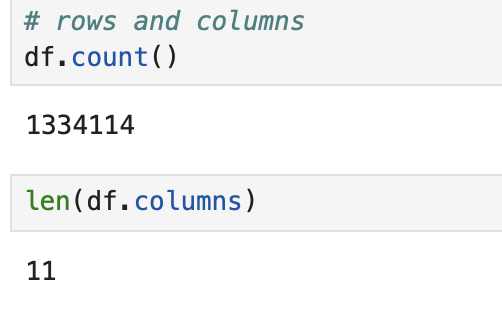


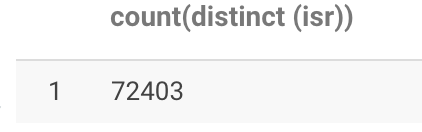
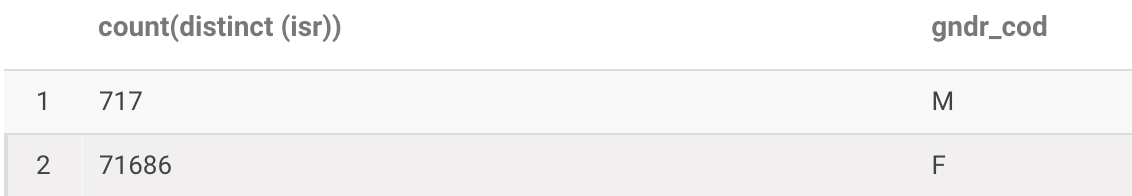
**Conclusion**

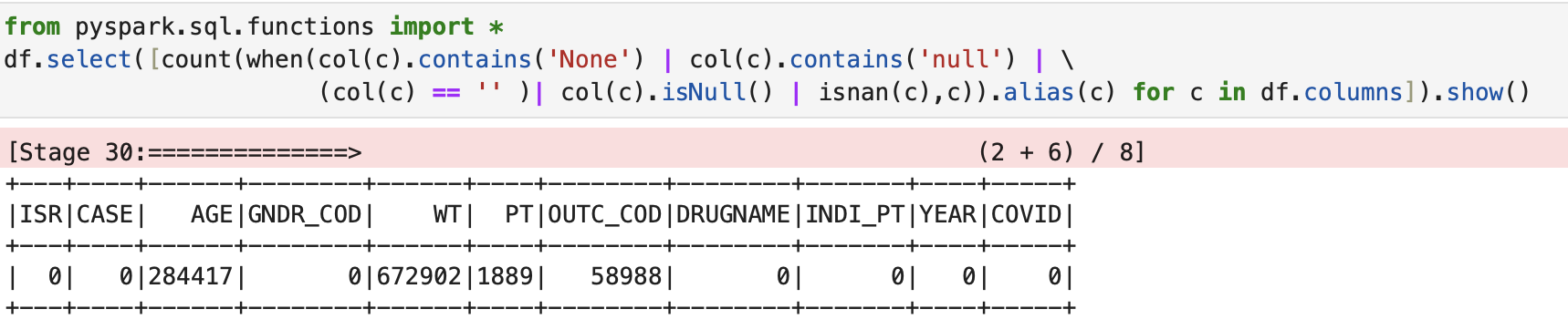
This study provides valuable insights into the treatment of metastatic breast cancer, focusing on the relationship between commonly used medications and their associated adverse reactions. Through a thorough analysis of a large and complex dataset, several important patterns were identified:

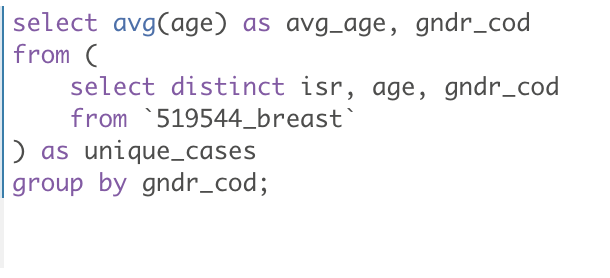
1. Key Findings:
   * Drugs such as IBRANCE and KISQALI are frequently prescribed, with symptoms like fatigue and neutropenia being common adverse reactions.
   * There is no significant difference in patient outcomes between pre- and post-COVID-19 cases.
2. Data:
   * Managing string-based features (e.g., symptoms and drug names) for predictive modeling posed significant challenges, affecting model accuracy.
   * Computational limitations restricted the use of more advanced machine learning models.
   * Developing better methods for handling categorical and string data can improve the accuracy of predictive models.

**Appendix**

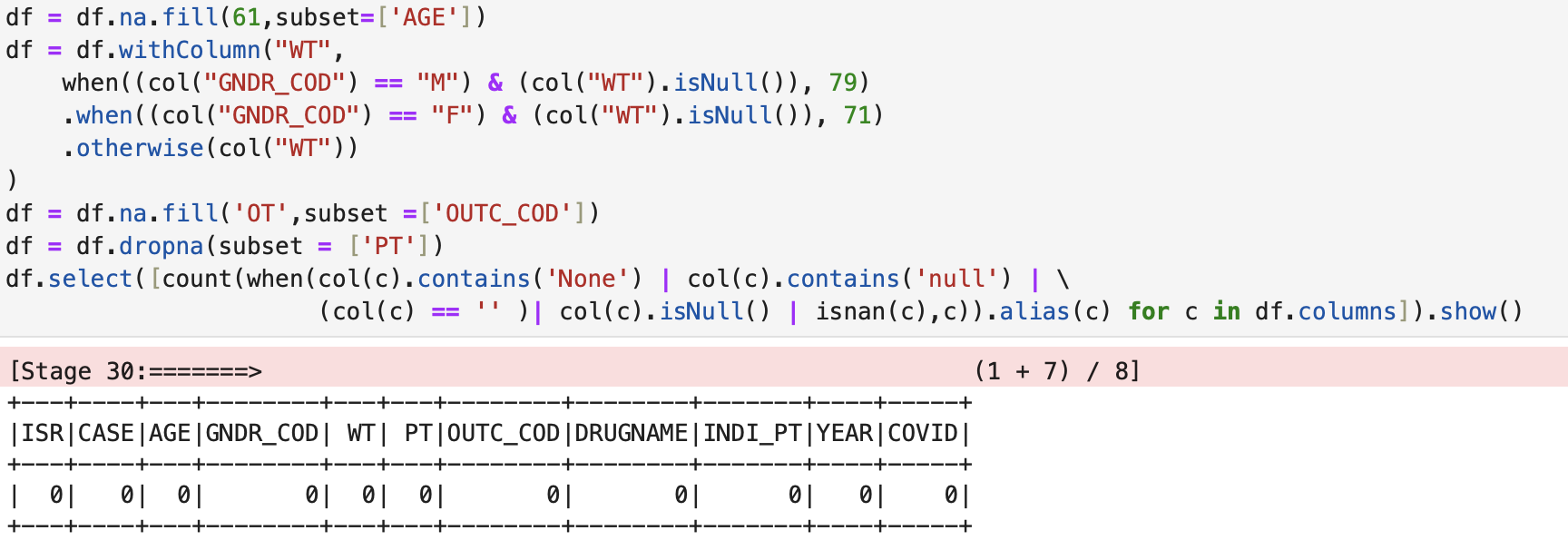
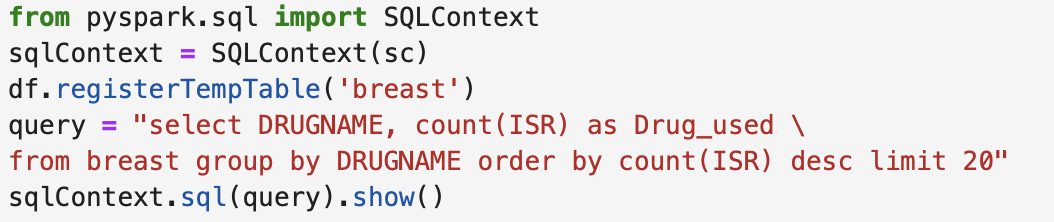
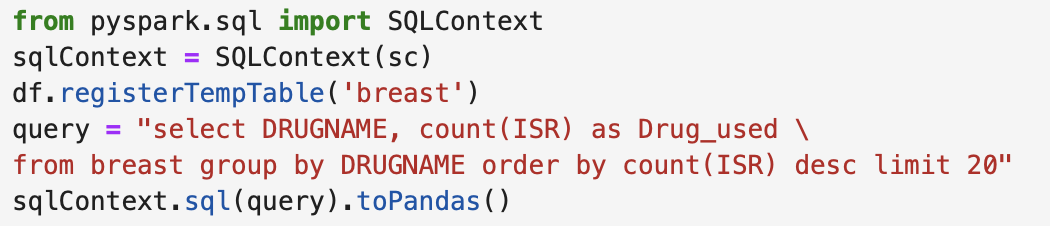
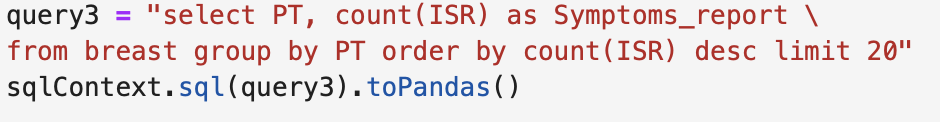
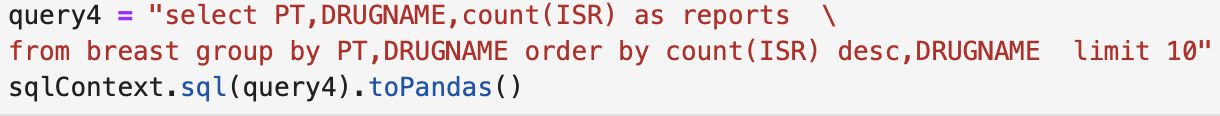
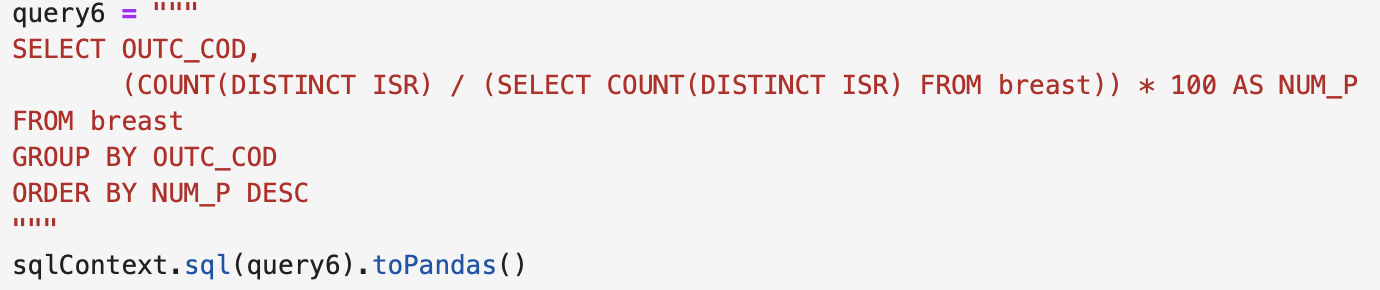


1. 
2. 



1. 



1. 
2. 
3. 
4. 
5. 
6. 
7. 