**Python Answer Key**

Read the File



1. How many rows and columns exist in the dataset?

*Code:* The shape method shows the dimensions of a dataset (number of rows by number of columns). This can be subset to show rows and columns.

**

*Answer:*

Number of rows in dataset: 198917

Number of columns in dataset: 14

1. What data type is each column?

*Code:* The dtypes attribute shows the data type for each column.



*Answer:*

Order int64

File\_Type object

SKU\_number int64

SoldFlag float64

SoldCount float64

MarketingType object

ReleaseNumber int64

New\_Release\_Flag int64

StrengthFactor float64

PriceReg float64

ReleaseYear int64

ItemCount int64

LowUserPrice float64

LowNetPrice float64

dtype: object

1. How many unique values are there in each column?

*Code:* The method nunique, which shows number of unique values, can be used with for loop to find unique values in each column



*Answer:*

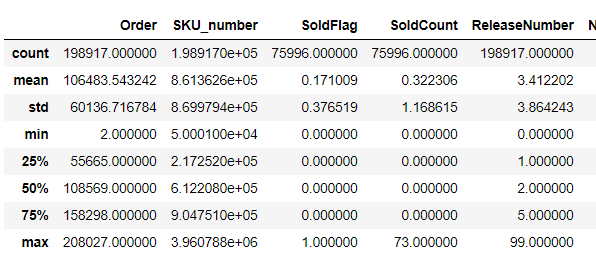
|  |  |
| --- | --- |
| Order | 198917 |
| File\_Type | 2 |
| SKU\_number | 133360 |
| SoldFlag | 2 |
| SoldCount | 37 |
| MarketingType | 2 |
| ReleaseNumber | 71 |
| New\_Release\_Flag | 2 |
| StrengthFactor | 197424 |
| PriceReg | 11627 |
| ReleaseYear | 85 |
| ItemCount | 501 |
| LowUserPrice | 12102 |
| LowNetPrice | 15403 |

1. Perform summary statistics for the entire dataset

*Code:* The describe method shows the summary statistics of numerical columns.



*Partial Output*:



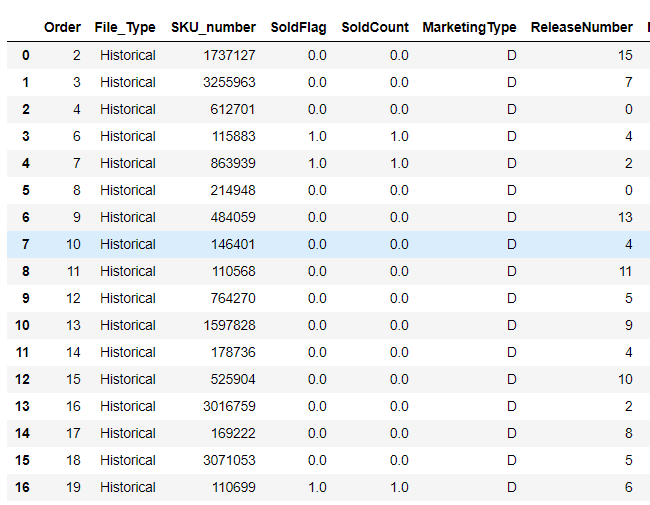
1. Display the first 17 rows in the dataset. In a separate line of code display the last 9 rows of data

The head and tail methods are useful to check the first and last rows of dataset respectively. The default displayed is 5 rows; however, this can be modified by specifying the desired number in the parenthesis.

*Code:*



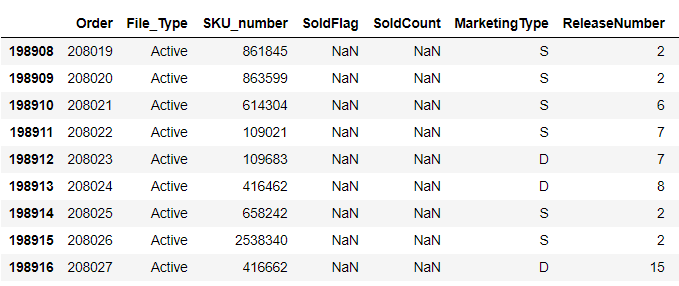
*Partial Output*:



*Code:*

**

*Partial Output*:



1. Are there any missing values in the data? What value is representing the missing value?

*Code*: The missing values can be checked using isna method of pandas dataframes and can be summed for each column.



*Answer*: Yes, there are missing values. Columns SoldFlag and SoldCount have 122,291 missing values each.

*Code*: The unique values for a column with missing values can be checked to see the value that represents missing value.



*Answer*: The missing values take nan values in the dataset based on the output below.



1. What is the % of records that are active sales records vs. historical sales records? (Hint: FileType column shows whether records are active or historical)

*Active Sales*: 61.80% records are active sales records.



*Historical Sales*: 38.20% records are historical sales records.



1. Create a new dataframe for Active sales only. Do the same for historical sales.

*Code*: The File\_Type column can be subset based on active and historical sales.



1. Rerun summary statistics for the new dataframes created above. Is there a huge difference between the two?

*Code*:



*Answer*:

The goal of this problem is to understand if there are any differences between active and historical sales data in terms of trends etc. This is a bit open ended. Ex: There are more orders in the active data set versus the historical dataset. On average, there were more new releases in the historical dataset. The avg Pricereg is higher in historical dataset and so is the itemcount. The main takeaway is that dataset specific summary statistics may inform modelbuilding/further analytics.

1. Among active sales, which SKU has the highest strength factor? Check the same for Historical sales.

*Code*: For active sales



*Answer*: SKU\_number: 3644308

*Code*: For historical sales



*Answer*: SKU\_number: 2259655

1. Which release year saw the lowest average user price?

*Code*:



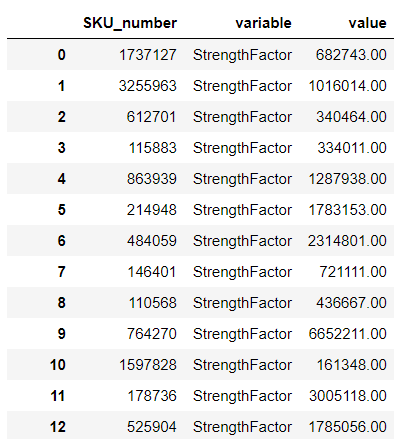
*Answer*: Release Year: 1923; Average User Price: 5.08

1. Reshape/ Pivot the data to create a dataframe that has SKUs in one column, all continuous numeric variables (StrengthFactor, PriceReg, ItemCount, LowUserPrice, LowNetPrice) as values in one column and all their respective numerical values in a 3rd column (Hint: this dataframe should have only 3 columns in total- SKU, Variable, Value).

*Code*: The melt method is like the pivot tables in excel or melt/cast in R.



*Answer*:

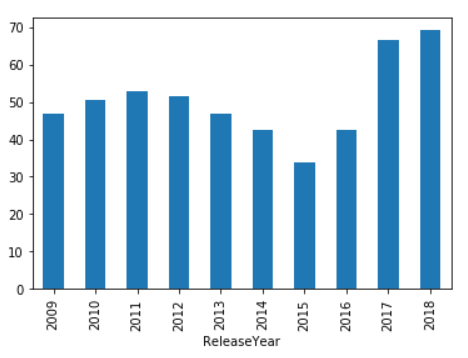


1. Create a chart/plot to show average lownetprice trend by release year for last 10 years.

*Code*: The charts in this answer keys are created using he matplotlib package but any other packages graph packages or base Python plots can be used.



*Answer*:

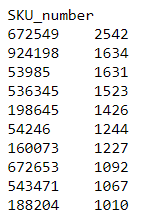


1. List the top 10 and bottom 10 SKUs by number of orders for both active and historical sales.

*Active Sales, Top 10 SKUs*:



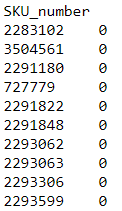
*Output*:



*Active Sales, Bottom 10 SKUs*:



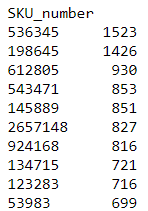
*Output*:

**

*Historical Sales, Top 10 SKUs*:



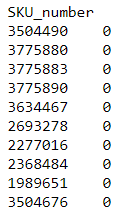
*Output*:



*Historical Sales, Bottom 10 SKUs*:



*Output*:



1. Analyze the relationship between StrengthFactor and PriceReg (Hint: a chart/plot might help). Create 3 charts in total – one for the entire data, and one each for historical and active sales.

*Code*:



*Plot (Just one shown for Entire Dataset)*:

