Summary of Adza Week 3 Subset.ipynb

The Jupyter notebook "Adza Week 3 Subset.ipynb" demonstrates basic data manipulation and preprocessing tasks using the Python library Pandas on big_mart_sales.csv. The dataset appears to contain sales data for a retail chain, with columns such as Item_Identifier, Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, Item_MRP, Outlet_Identifier, Outlet_Establishment_Year, Outlet_Size, Outlet_Location_Type, Outlet_Type, and Item_Outlet_Sales. The notebook focuses on loading, filtering, transforming, and handling missing values in this dataset.

The notebook begins by importing the Pandas library and loading the big_mart_sales.csv dataset into a DataFrame called info. The head() method is used to display the first five rows, revealing a mix of numerical and categorical data, including item details (e.g., weight, MRP, type) and outlet characteristics (e.g., size, location, type). A new DataFrame df is created as a copy of info for further manipulation.

Two subsets of the data are created:

1. **set1**: Filters rows where the Item_MRP (Maximum Retail Price) is less than 100. The resulting subset, displayed with head(), includes items like soft drinks, household goods, and frozen foods, with MRPs ranging from approximately 48 to 97.
2. **set2**: Filters rows where the Outlet_Location_Type is 'Tier 1' and the Outlet_Establishment_Year is either 1987, 1988, or 2009. However, the head() output shows an empty DataFrame, indicating no records match these criteria, possibly due to no outlets in Tier 1 being established in those specific years.

The notebook then performs a transformation on the Outlet_Type column by mapping categorical values to numerical codes using a dictionary (e.g., 'Supermarket Type1' to 1, 'Grocery Store' to 3). This mapping is applied to the Outlet_Type column in df, converting it from strings to integers, as verified by displaying the updated DataFrame with head().

Next, the notebook addresses missing values. A check using isna().sum() reveals 1,463 missing values in Item_Weight and 2,410 in Outlet_Size. To handle missing values in Outlet_Size, the notebook imputes them with the value 'Medium' using the loc method. A subsequent check confirms that Outlet_Size no longer has missing values, while Item_Weight still has 1,463. A commented-out line suggests an intention to impute missing Item_Weight values with 0, but this step was not executed.

Overall, the notebook illustrates fundamental Pandas operations, including data loading, filtering, mapping categorical values, and handling missing data. It serves as a practical exercise in data preprocessing, likely part of a learning module (Week 3) for data analysis or a related course. The empty result for set2 suggests a need to verify the dataset's contents or adjust the filtering criteria, and the unexecuted Item_Weight imputation indicates a potential area for further refinement.