4.4 Data_Wrangling_and_Subsetting

PREPARATION

Dropping a column

Reviewing a column

Renaming a column

Changing data type

Transposing data

Adding an index column

Creating a data dictionary

Creating a subset of the data frame

Indexing

TASK

Changing data type

Renaming columns

Find count of all values in a column

Using a data dictionary

Creating a subset

Subset of multiple items

Find information relating to one value

4.5 Data Consistency

Creating a new data frame

Looking for mixed data types

Changing data type

Looking for missing observations

Create a dataframe of the null values in one column

Create a data frame without the null values

Finding duplicate rows

New dataframe without duplicates

4.5 Data Consistency Exercise

Initial data checks

Changing data type

Check for mixed data

Look for missing observations

Create a new dataframe with null values removed

Find duplicate rows

4.6 Combine and Export

Importing and Initial Checks

Merging two dataframes

Export as pickle file

Export to Google Drive

4.6 Combine and Export Exercise

Notes on difficulty importing .pkl or .csv

Importing from Google Drive

Initial checks

Merging dataframes – Difficulties

Working with sample only

Merging dataframes

4.7 Define New Variables

# In this script, I find the values in a column. I categorise those values as new variable, and then create a new column for that new variable.

# E.g. busiest times of day

Defining and running a function

Importing libraries and drive

Import df and initial checks

Using a subset (first million rows)

Creating a new function - a price flag - using if/elif/else

Applying function - to a new 'price_range' column

Finding maximum value in a column

Creating the same function - a price flag - using loc()

Apply price flag using loc()

Counting all values in a column

Using if/elif/else to create a busy day flag – a NEW VARIABLE

Populating new column with created variable

TASK

Value counts for a column

Amending our newly created 'busiest_day' column

Changing column name

Putting new result into column

Creating a busiest times column

Find busiest times with .value_counts on a column

Create new variable 'result_hours' - using if/elif/else based on busiest times

Create new column and attach variable

4.8 Group and Aggregate

Applying values in a new column - this time having grouped by a column first

Import libraries, drive, df

Make a subset of the dataframe

Use Groupby - to summarise the data by a column

Applying an aggregate value to each value you have grouped by

Use loc() to create an if-else condition and apply to a new column

EXERCISE

Use loc() to create an if-else condition

Group by and check stats for those grouped values

Create a new column and define what goes in it

2nd version - Create a new column and define what goes in it

Reviewing NaN values

Renaming values

4.8 Group and Aggregate Practice

import pandas and a dataset

Group by and show stats

Renaming columns when showing the stats (not ‘min’ but ‘min_age’ etc)

Groupby more than one column

Import new data - setting data types

Groupby 2 columns ' 'sortby' count of another column

Create new column - use loc() and if-else to populate - with a default value

Use new column for a new groupby

View of the df with a groupbby applied

Another way - using loc() - to show the df with just one value from a column

4.9 Data Visualisation Intro

Loading libraries, drive, dataframe

Create bar chart

Sorting and making bar chart prettier

Saving bar chart

Creating histogram

Checking aggregated statistics

Creating scatterplot

Examining values of a column

More granular histogram

Creating a sample of data

Using just the sample

Creating a line chart

4.9 Data Vis PREP

Loading libraries and data frame

Changing column names

Change data type

Basic checks on data

Looking for missing values / observations

Excluding columns

Merging data frames

Troubleshooting when merging

Tidying new dataframe

Export to drive as pickle

4.9 Data Vis CHARTS

Instructions

Loading libraries, drive, dataframe

Create a bar chart

Save chart

Line chart - money spent at different times of day

Creating a sample

Smaller dataframe of just needed columns

Create line chart

Create bar chart - products per time of day

Bar chart of only top 25 products - using head()

Bar chart of top 25 products - at one specific time

Checking price of a specific product

Checking average price of all products

Line chart - age vs average no. dependents

Scatterplot - age vs income

Export chart

4.10 Create and Analyse Profiles_pt2

Import libraries and data

Initial checks

Creating my 3 subsets

Day 0 Subset

Office hours subset

South subset

Crosstab 1 - Day 0 profile

Creating a new column for High Spending New Customers

Visualising proportion of profile by region

Visualising proportion of profile by family circumstance

Office hours profile - comparing a profile with whole population

Sales by department in office hours v overall

Line chart: alcohol sales by age

Histogram: alcohol sales by age

Genders of people buying alcohol

Key stats - those buying alcohol

Visualise alcohol profile by region

PROFILE of the SOUTH

Creating a smaller data frame - based on total $ sales

Creating a smaller data frame - based on count of sales