Assignment_2_data_descriptions.ipynb_jhwa426
파일 수정 보기 삽입 런타임 도구 도움말 모든 변경사항이 저장됨
+ 코드 + 텍스트

Assignment 2

Welcome to assignment #2! In this assignment we're going to be doing some exploratory data analysis (EDA) on a dataset using pandas.

Note: You need to answer all questions in complete sentences. Failure to do so will result in part or zero marks.

Make sure that you can save this notebook to your own Google Drive space! Otherwise, your answers will not be saved!

Before you start, I'd like you to watch the following video. This is a good first introduction and review of some of the topics we'll be covering in this course.

```
[22] from IPython.display import YouTubeVideo
     YouTubeVideo("I10q6fjPxJ0")
```


```
[23] # First, load our libraries
     import pandas as pd
     import scipy.stats
     import numpy as np
     import pandas.api.types
     import math

     from IPython.core.magic import register_cell_magic
```

Load house price data

```
[24] hp = pd.read_csv("https://busan302.mycourses.work/data/house_price_train.csv") #1460 rows
     hp.head() # peek at first 5 lines!
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000

5 rows × 81 columns

Data dictionary

Here are brief explanations of our variables. You will **rarely** have a detailed dictionary in any organization. This means that your initial explorations of the data are absolutely crucial so that you can understand your data!

```
[25] from IPython.display import IFrame
     IFrame(src='https://busan302.mycourses.work/data/house_price_data_description_short.html', width=800, height=300)
```

Data fields

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits

Exploratory data analysis

Let's see how pre-1945 houses differ from houses built post-1980s.

Q1. Complete the code in the cells below. (2 marks total)

Replace the elipses '...' with appropriate goals to meet the defined goals.

```
[26] # Create subsets of old (pre-1945) and new houses (1980 onwards)

     #Old homes

     old_homes = hp[hp["YearBuilt"] < 1945]

     #New homes
     new_homes = hp[hp["YearBuilt"] >= 1980]
```

```
[27] # Find the means of the numeric variables in the new_homes and old_homes dataframes.
     #The means of New homes

     nh = np.mean(new_homes)

     #The means of Old homes

     oh = np.mean(old_homes)
```

```
/usr/local/lib/python3.7/dist-packages/numpy/core/fromnumeric.py:3438: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise
  return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

Compare new and old homes

The next cell will concatenate the averages you found for new and old homes. If you did the previous question correctly, you will see average values for new and old homes in the output.

Q2. Compare the rows that is generated from the code that follows. Answer each of the following:

(a) What general impressions do you have about new versus older homes?

(b) Briefly describe how newer and older homes differ.

(c) Do any of the differences surprise you?

(d) What do you think accounts for the differences?

```
[28] # Consider the data from this table
     df = pd.concat([oh, nh], axis=1) # Take the Series created from the previous question. Concat will place them side by side in a dataframe
     df.columns = ['old houses', 'new houses'] # Give the columns descriptive names
     df # jupyter prints out the last line for us
```

	old houses	new houses
Id	739.682482	717.622549
MSSubClass	61.478102	60.751634
LotFrontage	62.796000	72.641326
LotArea	8977.890511	10258.459150
OverallQual	5.368613	7.163399
OverallCond	6.215328	5.116013
YearBuilt	1922.470803	2000.171569
YearRemodAdd	1971.463504	2000.928105
MasVnrArea	12.178832	157.677686
BsmtFinSF1	188.273723	526.535948
BsmtFinSF2	18.496350	22.266340
BsmtUnfSF	594.281022	694.290850
TotalBsmtSF	801.051095	1243.093137
1stFlrSF	952.708029	1268.500000
2ndFlrSF	479.817518	440.457516
LowQualFinSF	27.478102	0.392157
GrLivArea	1460.003650	1709.349673
BsmtFullBath	0.200730	0.508170
BsmtHalfBath	0.036496	0.027778
FullBath	1.299270	1.911765
HalfBath	0.222628	0.517974
BedroomAbvGr	2.930657	2.802288
KitchenAbvGr	1.087591	1.004902
TotRmsAbvGrd	6.554745	6.862745
Fireplaces	0.503650	0.697712
GarageYrBlt	1948.304348	2000.377483
GarageCars	1.226277	2.233660
GarageArea	326.562044	586.947712
WoodDeckSF	50.529197	126.401961
OpenPorchSF	34.689781	65.282680
EnclosedPorch	69.726277	5.691176

A2. Write your answers in this cell

(a) Generally, new homes' floor space has bigger space than older homes.

(b) The most significant difference between newer and older homes is the average sale price of new homes are approximately $100,000 higher than old homes and newer homes show that there is a high number of trading.

(c) The surprise is that the average floor space of the new houses is much bigger than the old ones, which means the new houses are supposed to be built with large bedrooms than other area. Also, there is no difference in annual home sales and monthly home sales between newer and older homes.

(d) The differences show that the construction trend has been developed. The floor space of the old homes is built with plenty of small bedrooms; however, newer homes tend to be built on a large scale of bedrooms.

Looking at differences

We'd like to compare the differences between older and newer houses a little more rigorously. Let's use a t-test to statistically compare the average value of new homes and old homes across a number of characteristics.

The loop below looks over the columns from our dataframes. If the column is numeric it runs a t-test. The output is a printout of the variable that is used in the t-test and the difference that results when we substrat the average for new homes minus the average for old homes.

For example, look at the line:

```
LotArea       (new house value - old house value) = 1280.57          p-value=0.0009
```

This output means that the average lot area of new homes is 1280.6 square feet larger than the average lot area of older homes. The p-value is the likelihood that we would see this difference if the null hypothesis was true. That means there's a 0.0009% likelihood of seeing this difference under the null! i.e., not very likely! :-)

Q3. The presenter in the video talked about two types of data to conduct a t-test: categorical and numeric variables (2 marks).

(a) What is the null hypothesis of the t-test?

(b) Give an example of the categorical and numeric variables we are considering when we conduct the t-test in the code below.

A3. Place your answers below.

(a) The null hypothesis is the p-value equivalent to 0.

(b) Example variables are:

- categorical: Age group, Gender
- numeric: Weight, Height

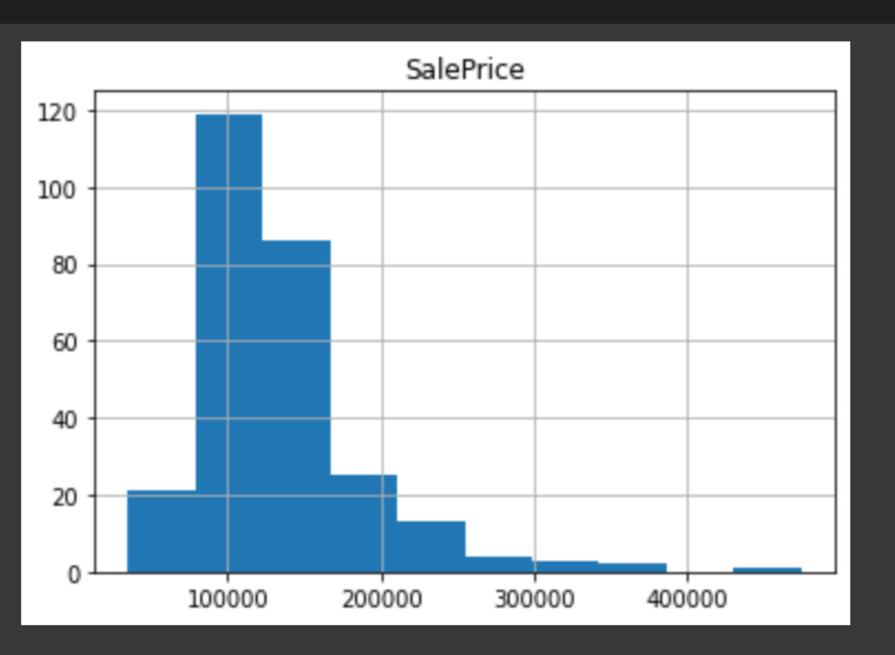
```
[29] for var in old_homes.columns:

         # Next line checks that the dataframe column is numeric.
         if not pandas.api.types.is_numeric_dtype( old_homes[var].dtype ):
             continue # skip all non-numeric columns

         # t-test: do a t-test to compare means for the column in the old and new house dataframes
         t = scipy.stats.ttest_ind( old_homes[var], new_homes[var] )

         if t[1] < 0.05: # Only print cases where the p-value is less than 5%. We'll say for our purposes this is significant.
             # The f"" syntax is a formatted string. It may be a little cryptic if you're new to python!
             print( f"{var} \t (new house value - old house value) = {round( new_homes[var].mean() - old_homes[var].mean(), 2)} \tp-value={round(t[1],4)}" )
```

```
LotArea          (new house value - old house value) = 1280.57  p-value=0.0009
OverallQual      (new house value - old house value) = 1.79     p-value=0.0
OverallCond      (new house value - old house value) = -1.1     p-value=0.0
YearBuilt        (new house value - old house value) = 77.7     p-value=0.0
YearRemodAdd     (new house value - old house value) = 29.46    p-value=0.0
BsmtFinSF1       (new house value - old house value) = 338.26   p-value=0.0
BsmtUnfSF        (new house value - old house value) = 100.01   p-value=0.0029
TotalBsmtSF      (new house value - old house value) = 442.04   p-value=0.0
1stFlrSF         (new house value - old house value) = 315.79   p-value=0.0
LowQualFinSF     (new house value - old house value) = -27.09   p-value=0.0
GrLivArea        (new house value - old house value) = 249.35   p-value=0.0
BsmtFullBath     (new house value - old house value) = 0.31     p-value=0.0
FullBath         (new house value - old house value) = 0.61     p-value=0.0
HalfBath         (new house value - old house value) = 0.3      p-value=0.0
BedroomAbvGr     (new house value - old house value) = -0.13    p-value=0.031
KitchenAbvGr     (new house value - old house value) = -0.08    p-value=0.0
TotRmsAbvGrd     (new house value - old house value) = 0.31     p-value=0.0103
Fireplaces       (new house value - old house value) = 0.19     p-value=0.0
GarageCars       (new house value - old house value) = 1.01     p-value=0.0
GarageArea       (new house value - old house value) = 260.39   p-value=0.0
WoodDeckSF       (new house value - old house value) = 75.87    p-value=0.0
OpenPorchSF      (new house value - old house value) = 30.59    p-value=0.0
EnclosedPorch    (new house value - old house value) = -64.04   p-value=0.0
MiscVal          (new house value - old house value) = -41.75   p-value=0.0024
SalePrice        (new house value - old house value) = 100085.04        p-value=0.0
```

Q4. Plot histograms for the SalePrice of older homes and newer homes. Describe the shape of each histogram (3 marks).

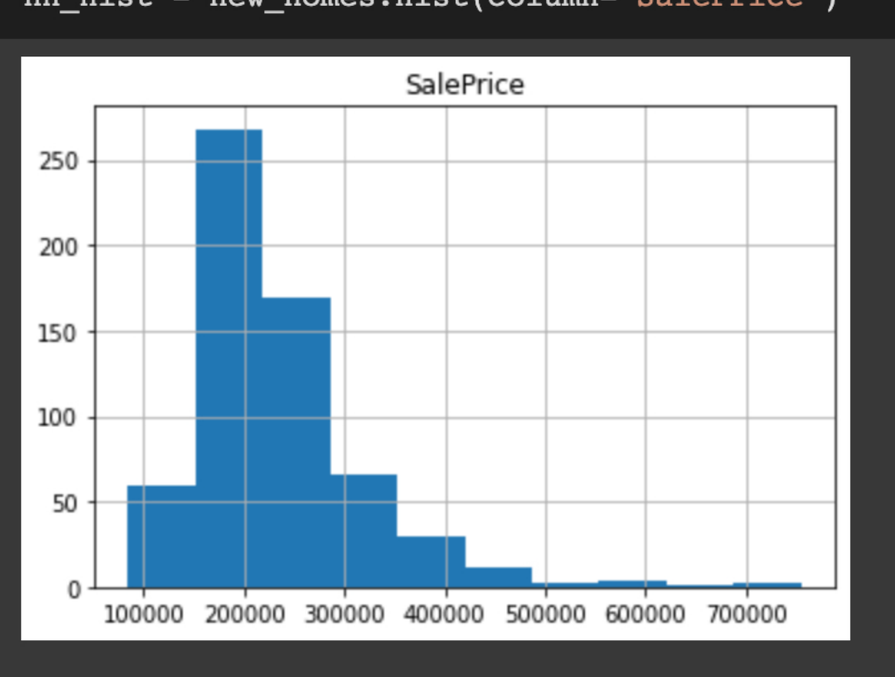
You can find out more about how to plot histograms in pandas here.

```
[30] # older homes histogram

     oh_hist = old_homes.hist(column="SalePrice")
```



A4. I would describe the distribution as a right-skewed plot of the SalePrice of older homes. About 120 houses have been sold for approximately 100,000 dollars and the highest sale price of house was approximately more than 400,000 dollars.

```
[31] # newer homes histogram

     nh_hist = new_homes.hist(column="SalePrice")
```



A4. I would describe the distribution as a right-skewed plot of the SalePrice of newer homes. About more than 270 houses have been sold for approximately 200,000 dollars and the highest sale price of the house was approximately more than 700,000 dollars.

Use the original dataframe (hp) to answer the following questions.

Q5. (a) When people are selling their home they look at many different aspects of a house. Choose one variable that you think is important when buying a home and explain your reasoning (1 mark).

Q5 (b) Show a scatter plot of your variable of choice versus SalePrice (1 mark).

Q5 (c) Explain the relationship you see. (1 mark)

```
[32] # The top 10 of the highest house sale price for reference (Optional - The sorted highest price helps to decide to sell homes)

     hp["SalePrice"].sort_values(ascending=False).head(10)
```

```
691     755000
1182    745000
1169    625000
898     611657
803     582933
1046    556581
440     555000
769     538000
178     501837
798     485000
Name: SalePrice, dtype: int64
```

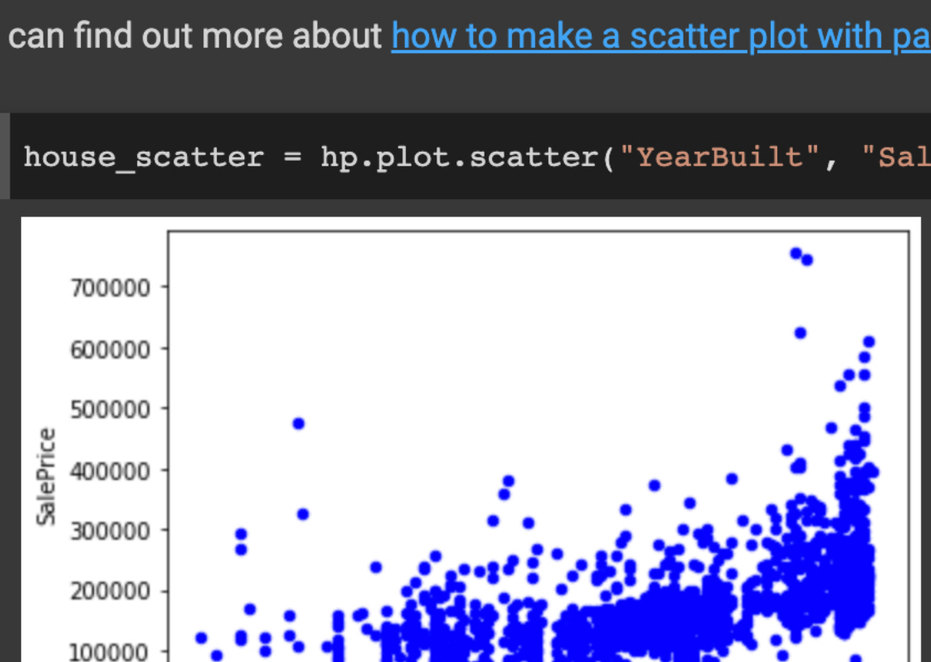
A5 (a) I think this variable is important ...

- The most significant variable is the house's sale price because people wish to sell at a reasonable home price, so the selling price will play a significant role in whether or not people should sell a home.

A5 (b) Show the scatter plot below.

You can find out more about how to make a scatter plot with pandas here. The syntax is [dataframe].plot.scatter()

```
house_scatter = hp.plot.scatter("YearBuilt", "SalePrice", c="blue")
```



A5 (c) I see the following relationship:

- The scatter plot shows a positive relationship that the sale price of homes gradually an increasing since 1900. Also, house trading has noticeably increased since 1920 regardless of the year of the houses.

0초 오후 9:01에 완료됨