

Code & Output

Final Project

Code:

```
# Julia Cuellar
# DSC 550
# Final project

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Display pizza place data
def read_file():
    pizza = pd.read_csv('pizzaplace.csv')
    print('Original pizza data:\n', pizza)

# Display described, summarized, and length of pizza place data
def des_sum_len():
    pizza = pd.read_csv('pizzaplace.csv')
    print('Described pizza data:\n', pizza.describe())
    print('Summarized pizza data:\n', pizza.describe(include=['O']))
    print('Length of pizza data:\n', len(pizza))

# Display bar chart of pizza name
def showBar_Pname():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza['name'].value_counts().plot(kind='barh').invert_yaxis()
    plt.title('Pizza name')
    plt.show()

# Display bar chart of pizza size
def showBar_Psize():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza['size'].value_counts().plot(kind='barh')
    plt.title('Pizza size')
    plt.show()

# Display pie chart of pizza type
def showPie_Ptype():
    pizza = pd.read_csv('pizzaplace.csv')
    plt.pie(pizza['type'].value_counts(), autopct=lambda p: f'{p:.2f}%',
    labels=['classic', 'supreme', 'veggie',
    'chicken'])
    plt.title('Pizza type')
    plt.show()
```

```

# Display boxplot of pizza price
def showBoxplot_Pprice():
    pizza = pd.read_csv('pizzaplace.csv')
    sns.boxplot(pizza['price'])
    plt.title('Pizza price')
    plt.show()

# Check the nulls from pizza file
def check_null():
    pizza = pd.read_csv('pizzaplace.csv')
    print("Display pizza data with null:\n", pizza.isnull())
    print("Display counts of null from pizza data:\n", pizza.isnull().sum())

# Rename unname column then drop along with id and date
def rename_drop():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    print('Pizza data updated:\n', pizza.head(5))

# Check for outlier in pizza size column by counts
def size_count():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    print('Pizza size count:\n', pizza['size'].value_counts())

# Check for outlier in pizza price column by describe then remove and update
def price_out_r_up():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    print('Pizza price:\n', pizza['price'].describe())
    p_price = pizza[pizza['price'] >= 35].index
    pizza.drop(p_price, inplace=True)
    print('Described pizza price:\n', pizza['price'].describe())
    sns.boxplot(pizza['price'])
    plt.title('Pizza price updated')
    plt.show()

# Redisplay pizza place data with described, summarized, and length
def pizza_up():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    p_price = pizza[pizza['price'] >= 35].index
    pizza.drop(p_price, inplace=True)
    pizza.to_csv('pizza.csv')
    print('Pizza data updated:\n', pizza)
    print('Described pizza data updated:\n', pizza.describe())

```

```

print('Summarized pizza data updated:\n', pizza.describe(include=['O']))
print('Length of pizza data updated:\n', len(pizza))

# Create a multiple linear regression model for size of pizza vs type of
pizza purchased
def reg_model_svt():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    p_price = pizza[pizza['price'] >= 35].index
    pizza.drop(p_price, inplace=True)
    fit = ols('price ~ C(size) + C(type)', data=pizza).fit()
    print("Multiple linear regression model for size of pizza vs type of
pizza purchased:\n", fit.summary())
    res = fit.resid
    fig = sm.qqplot(res, fit=True, line="45")
    plt.title('Multiple linear regression plot')
    plt.show()

# Display frequency table for size of pizza vs type of pizza purchased
def showFT_svt():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    p_price = pizza[pizza['price'] >= 35].index
    pizza.drop(p_price, inplace=True)
    pizza_svt = pd.crosstab(pizza['size'], pizza['type'])
    print("Cross table of size of pizza vs type of pizza purchased:\n",
pizza_svt)

# Create a simple linear regression model for name of pizza vs price of pizza
purchased
def reg_model_nvp():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    p_price = pizza[pizza['price'] >= 35].index
    pizza.drop(p_price, inplace=True)
    fit = ols('price ~ C(name)', data=pizza).fit()
    print("Simple linear regression model for name of pizza vs price of pizza
purchased:\n", fit.summary())
    res = fit.resid
    fig = sm.qqplot(res, fit=True, line="45")
    plt.title('Simple linear regression plot')
    plt.show()

# Display plot for name of pizza vs type of pizza purchased
def showPlot_nvp():
    pizza = pd.read_csv('pizzaplace.csv')
    pizza.rename(columns={'Unnamed: 0': 'num'}, inplace=True)
    pizza.drop(['num', 'id', 'date'], axis=1, inplace=True)
    p_price = pizza[pizza['price'] >= 35].index
    pizza.drop(p_price, inplace=True)

```

```

sns.catplot(x='price', y='name', data=pizza)
plt.title('Price vs Name')
plt.show()
pizza_nvp = pd.crosstab(pizza['price'], pizza['name'])
print("Cross table of price of pizza vs name of pizza purchased:\n",
pizza_nvp)
pizza_name = pizza.groupby('name').count()
print("Display count of pizza name:\n", pizza_name)
pizza_price = pizza.groupby('price').count()
print("Display count of pizza price:\n", pizza_price)

if __name__ == "__main__":
    read_file()
    des_sum_len()
    showBar_Pname()
    showBar_Psize()
    showPie_Ptype()
    showBoxplot_Pprice()
    check_null()
    rename_drop()
    size_count()
    price_out_r_up()
    pizza_up()
    reg_model_svt()
    showFT_svt()
    reg_model_nvp()
    showPlot_nvp()

```

Output:

Original pizza data:

	Unnamed: 0	id	date	...	size	type	price
0	1	2015-000001	2015-01-01	...	M	classic	13.25
1	2	2015-000002	2015-01-01	...	M	classic	16.00
2	3	2015-000002	2015-01-01	...	M	veggie	16.00
3	4	2015-000002	2015-01-01	...	L	chicken	20.75
4	5	2015-000002	2015-01-01	...	L	veggie	18.50
...
49569	49570	2015-021348	2015-12-31	...	L	veggie	17.95
49570	49571	2015-021348	2015-12-31	...	S	classic	12.00
49571	49572	2015-021348	2015-12-31	...	M	chicken	16.75
49572	49573	2015-021349	2015-12-31	...	L	veggie	20.25
49573	49574	2015-021350	2015-12-31	...	S	chicken	12.75

[49574 rows x 8 columns]

Described pizza data:

	Unnamed: 0	price
count	49574.000000	49574.000000
mean	24787.500000	16.497762
std	14310.925459	3.621954
min	1.000000	9.750000
25%	12394.250000	12.750000
50%	24787.500000	16.500000
75%	37180.750000	20.250000
max	49574.000000	35.950000

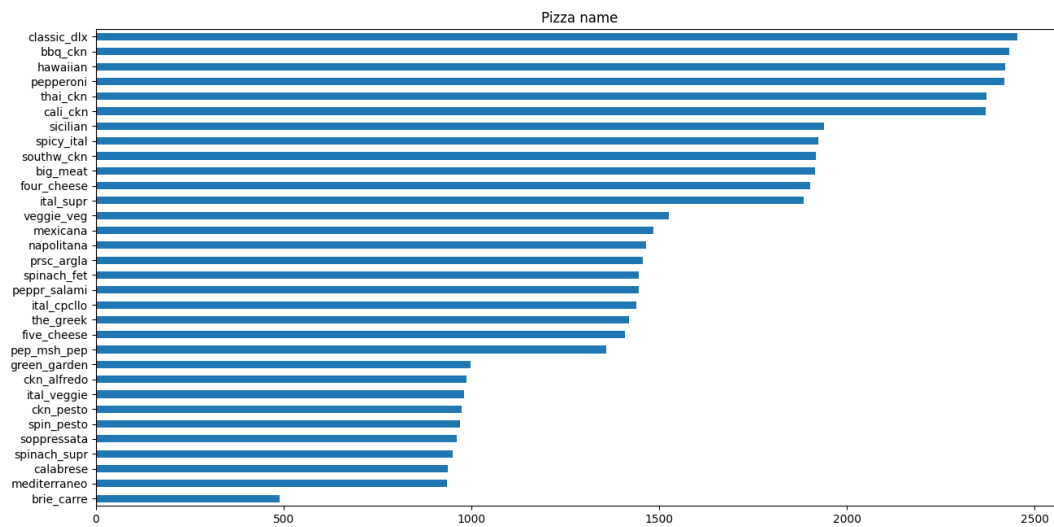
Summarized pizza data:

id	date	time	name	size	type
----	------	------	------	------	------

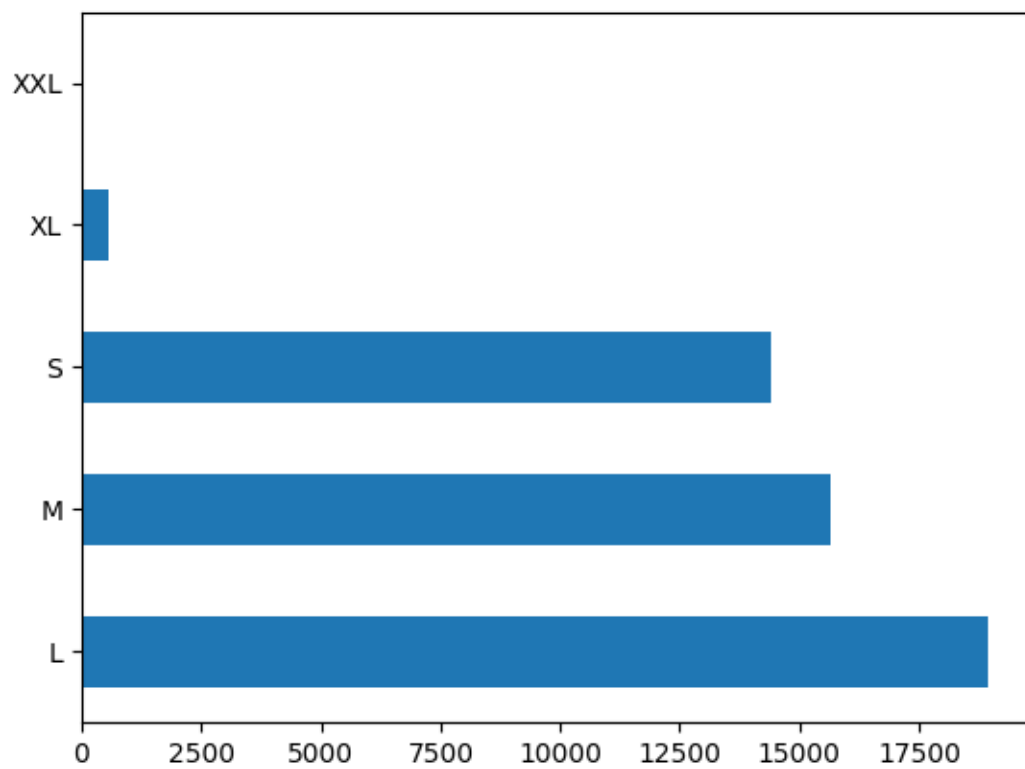
count	49574	49574	49574	49574	49574	49574
unique	21350	358	16382	32	5	4
top	2015-018845	2015-11-26	12:25:12	classic_dlx	L	classic
freq	28	266	28	2453	18956	14888

Length of pizza data:

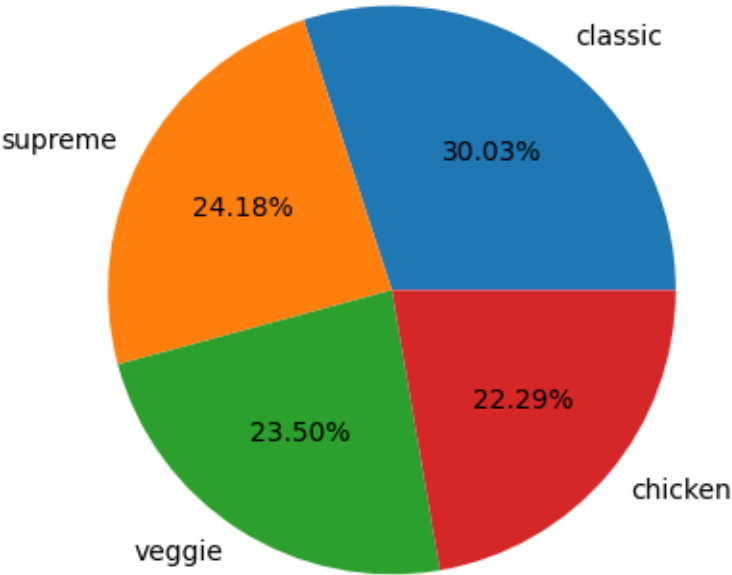
49574

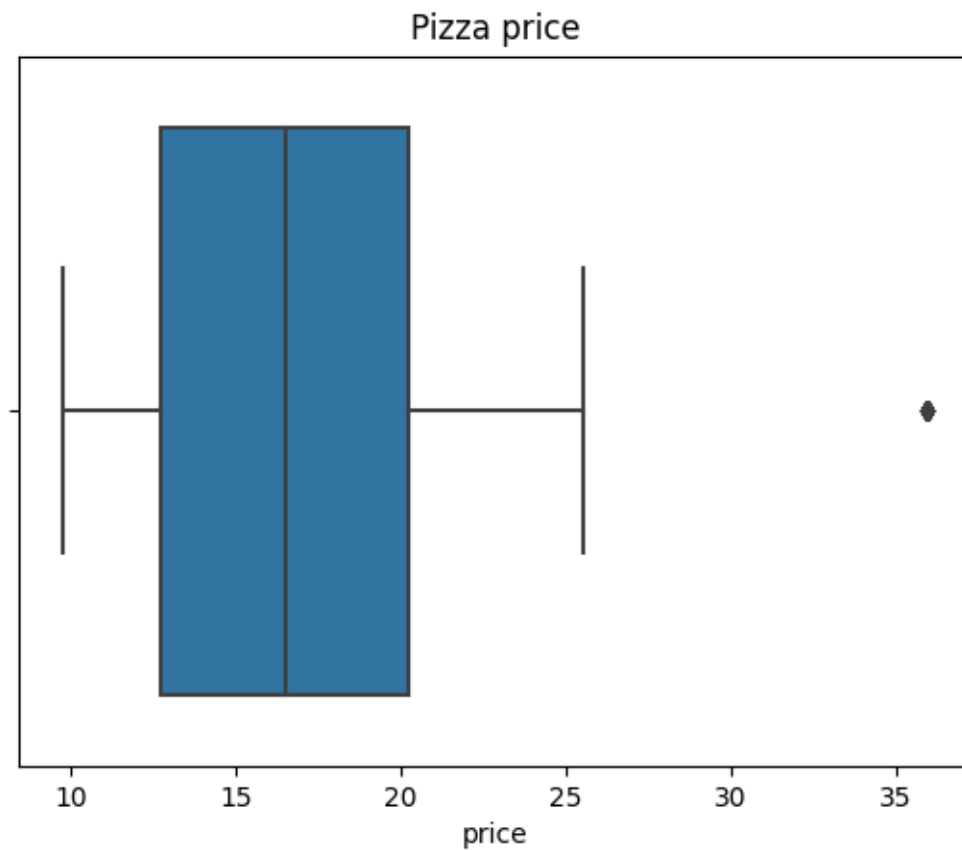


Pizza size



Pizza type





Display pizza data with null:

```

      Unnamed: 0   id  date  time  name  size  type  price
0      False False False False False False False False
1      False False False False False False False False
2      False False False False False False False False
3      False False False False False False False False
4      False False False False False False False False
...      ...  ...  ...  ...  ...  ...  ...
49569   False False False False False False False False
49570   False False False False False False False False
49571   False False False False False False False False
49572   False False False False False False False False
49573   False False False False False False False False

```

[49574 rows x 8 columns]

Display counts of null from pizza data:

Unnamed: 0 0

id 0

date 0

time 0

name 0

size 0

type 0

price 0

dtype: int64

Pizza data updated:

	time	name	size	type	price
0	11:38:36	hawaiian	M	classic	13.25
1	11:57:40	classic_dlx	M	classic	16.00
2	11:57:40	mexicana	M	veggie	16.00
3	11:57:40	thai_ckn	L	chicken	20.75
4	11:57:40	five_cheese	L	veggie	18.50

Pizza size count:

L 18956

M 15635

S 14403

XL 552

XXL 28

Name: size, dtype: int64

Pizza price:

count 49574.000000

mean 16.497762

std 3.621954

min 9.750000

25% 12.750000

50% 16.500000

75% 20.250000

max 35.950000

Name: price, dtype: float64

Described pizza price:

count 49546.000000

mean 16.486769

std 3.593327

min 9.750000

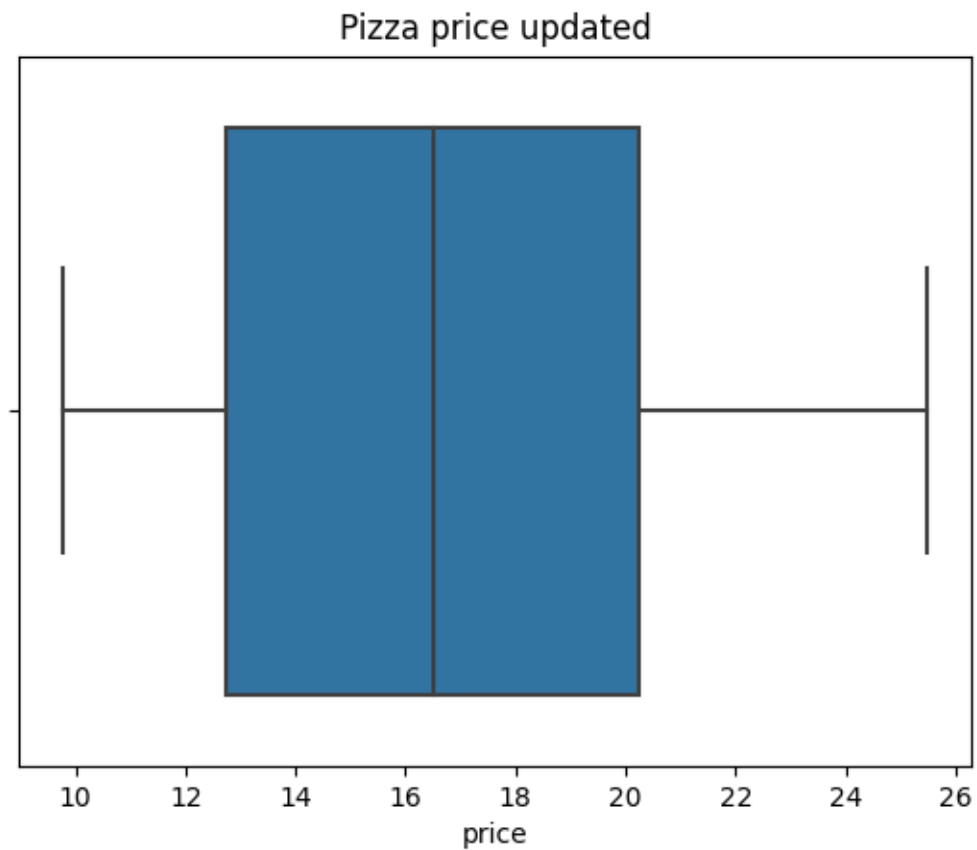
25% 12.750000

50% 16.500000

75% 20.250000

max 25.500000

Name: price, dtype: float64



Pizza data updated:

	time	name	size	type	price
0	11:38:36	hawaiian	M	classic	13.25
1	11:57:40	classic_dlx	M	classic	16.00
2	11:57:40	mexicana	M	veggie	16.00
3	11:57:40	thai_ckn	L	chicken	20.75
4	11:57:40	five_cheese	L	veggie	18.50
...
49569	21:23:10	four_cheese	L	veggie	17.95
49570	21:23:10	napolitana	S	classic	12.00
49571	21:23:10	ckn_alfredo	M	chicken	16.75
49572	22:09:54	mexicana	L	veggie	20.25
49573	23:02:05	bbq_ckn	S	chicken	12.75

[49546 rows x 5 columns]

Described pizza data updated:

```
      price
count 49546.000000
mean   16.486769
std     3.593327
min     9.750000
25%    12.750000
50%    16.500000
75%    20.250000
max    25.500000
```

Summarized pizza data updated:

```
      time      name  size  type
count  49546      49546 49546 49546
unique  16378       32    4     4
top    12:25:12 classic_dlx  L  classic
freq     28      2453 18956 14860
```

Length of pizza data updated:

49546

Multiple linear regression model for size of pizza vs type of pizza purchased:

OLS Regression Results

```
=====
=====
```

Dep. Variable:	price	R-squared:	0.847
Model:	OLS	Adj. R-squared:	0.847
Method:	Least Squares	F-statistic:	4.570e+04
Date:	Sat, 06 Mar 2021	Prob (F-statistic):	0.00
Time:	15:24:58	Log-Likelihood:	-87171.

No. Observations: 49546 AIC: 1.744e+05
Df Residuals: 49539 BIC: 1.744e+05
Df Model: 6
Covariance Type: nonrobust

=====

	coef	std err	t	P> t	[0.025	0.975]

Intercept	20.5052	0.015	1343.493	0.000	20.475	20.535
C(size)[T.M]	-3.8127	0.015	-250.543	0.000	-3.843	-3.783
C(size)[T.S]	-7.1114	0.016	-448.156	0.000	-7.143	-7.080
C(size)[T.XL]	7.0274	0.062	113.857	0.000	6.906	7.148
C(type)[T.classic]	-2.0326	0.018	-112.266	0.000	-2.068	-1.997
C(type)[T.supreme]	0.1538	0.019	8.277	0.000	0.117	0.190
C(type)[T.veggie]	-1.0795	0.019	-57.800	0.000	-1.116	-1.043

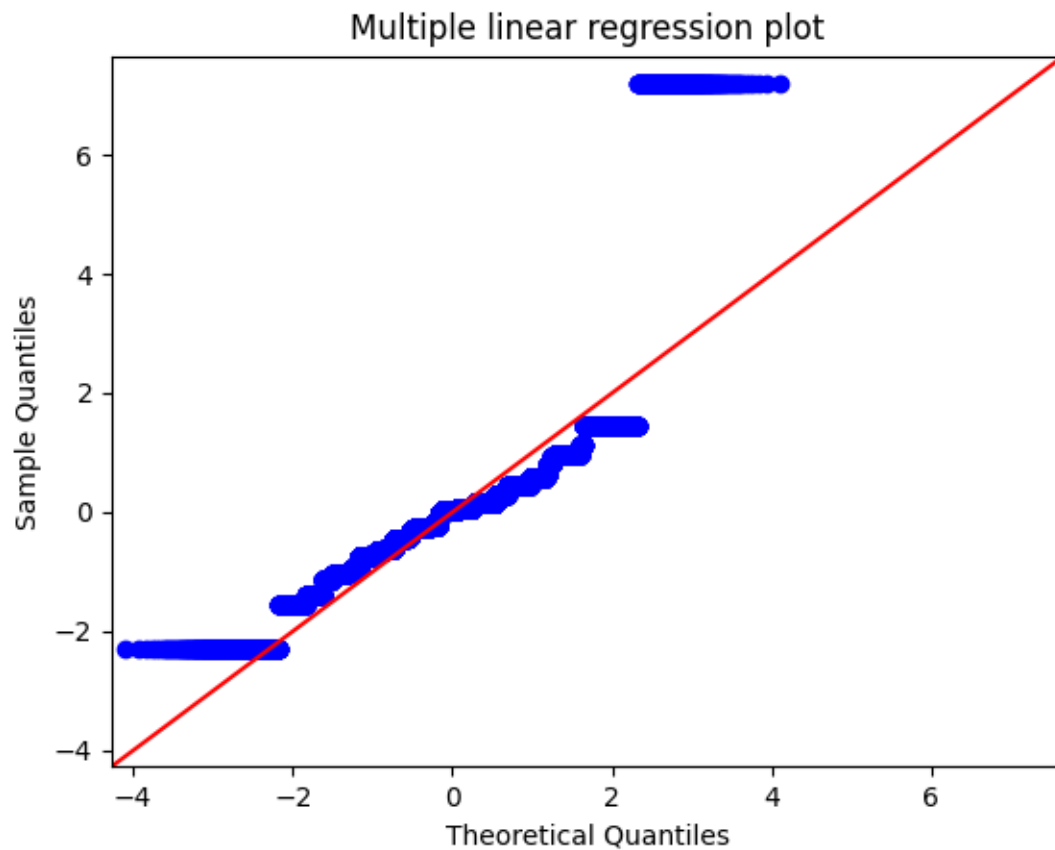
=====

Omnibus: 38426.463 Durbin-Watson: 2.006
Prob(Omnibus): 0.000 Jarque-Bera (JB): 1319807.442
Skew: 3.452 Prob(JB): 0.00
Kurtosis: 27.324 Cond. No. 11.7

=====

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



Cross table of size of pizza vs type of pizza purchased:

type chicken classic supreme veggie

size

L	4932	4057	4564	5403
M	3894	4112	4046	3583
S	2224	6139	3377	2663
XL	0	552	0	0

Simple linear regression model for name of pizza vs price of pizza purchased:

OLS Regression Results

=====

Dep. Variable:	price	R-squared:	0.315
Model:	OLS	Adj. R-squared:	0.314

Method: Least Squares F-statistic: 733.8
Date: Sat, 06 Mar 2021 Prob (F-statistic): 0.00
Time: 15:25:02 Log-Likelihood: -1.2431e+05
No. Observations: 49546 AIC: 2.487e+05
Df Residuals: 49514 BIC: 2.490e+05
Df Model: 31
Covariance Type: nonrobust

=====

	coef	std err	t	P> t	[0.025	0.975]
Intercept	17.5855	0.060	291.468	0.000	17.467	17.704
C(name)[T.big_meat]	-5.5855	0.091	-61.436	0.000	-5.764	-5.407
C(name)[T.brie_carre]	6.0645	0.147	41.161	0.000	5.776	6.353
C(name)[T.calabrese]	-0.5799	0.114	-5.069	0.000	-0.804	-0.356
C(name)[T.cali_ckn]	-0.1132	0.086	-1.318	0.188	-0.281	0.055
C(name)[T.ckn_alfredo]	-0.4627	0.112	-4.120	0.000	-0.683	-0.243
C(name)[T.ckn_pesto]	-0.4203	0.113	-3.724	0.000	-0.642	-0.199
C(name)[T.classic_dlx]	-2.0207	0.085	-23.733	0.000	-2.188	-1.854
C(name)[T.five_cheese]	0.9145	0.100	9.180	0.000	0.719	1.110
C(name)[T.four_cheese]	-0.6214	0.091	-6.823	0.000	-0.800	-0.443
C(name)[T.green_garden]	-3.5878	0.112	-32.065	0.000	-3.807	-3.368
C(name)[T.hawaiian]	-4.2605	0.085	-49.881	0.000	-4.428	-4.093
C(name)[T.ital_cpello]	-0.1349	0.099	-1.363	0.173	-0.329	0.059
C(name)[T.ital_supr]	0.1834	0.091	2.009	0.045	0.004	0.362
C(name)[T.ital_veggie]	-1.2560	0.113	-11.161	0.000	-1.477	-1.035
C(name)[T.mediterraneo]	-1.1396	0.115	-9.950	0.000	-1.364	-0.915
C(name)[T.mexicana]	0.4608	0.098	4.702	0.000	0.269	0.653
C(name)[T.napolitana]	-1.1327	0.098	-11.508	0.000	-1.326	-0.940

C(name)[T.pep_msh_pep]	-3.7264	0.101	-36.980	0.000	-3.924	-3.529
C(name)[T.pepperoni]	-5.1117	0.085	-59.821	0.000	-5.279	-4.944
C(name)[T.peppr_salami]	0.0694	0.099	0.702	0.483	-0.124	0.263
C(name)[T.prsc_argla]	-0.9807	0.099	-9.949	0.000	-1.174	-0.787
C(name)[T.sicilian]	-1.6204	0.091	-17.885	0.000	-1.798	-1.443
C(name)[T.soppressata]	-0.4932	0.113	-4.350	0.000	-0.715	-0.271
C(name)[T.southw_ckn]	0.5187	0.091	5.707	0.000	0.341	0.697
C(name)[T.spicy_ital]	0.5180	0.091	5.706	0.000	0.340	0.696
C(name)[T.spin_pesto]	-1.5072	0.113	-13.339	0.000	-1.729	-1.286
C(name)[T.spinach_fet]	-1.4920	0.099	-15.100	0.000	-1.686	-1.298
C(name)[T.spinach_supr]	-1.5037	0.114	-13.209	0.000	-1.727	-1.281
C(name)[T.thai_ckn]	0.7334	0.086	8.541	0.000	0.565	0.902
C(name)[T.the_greek]	2.1325	0.100	21.325	0.000	1.937	2.329
C(name)[T.veggie_veg]	-1.6126	0.097	-16.596	0.000	-1.803	-1.422

=====

=====

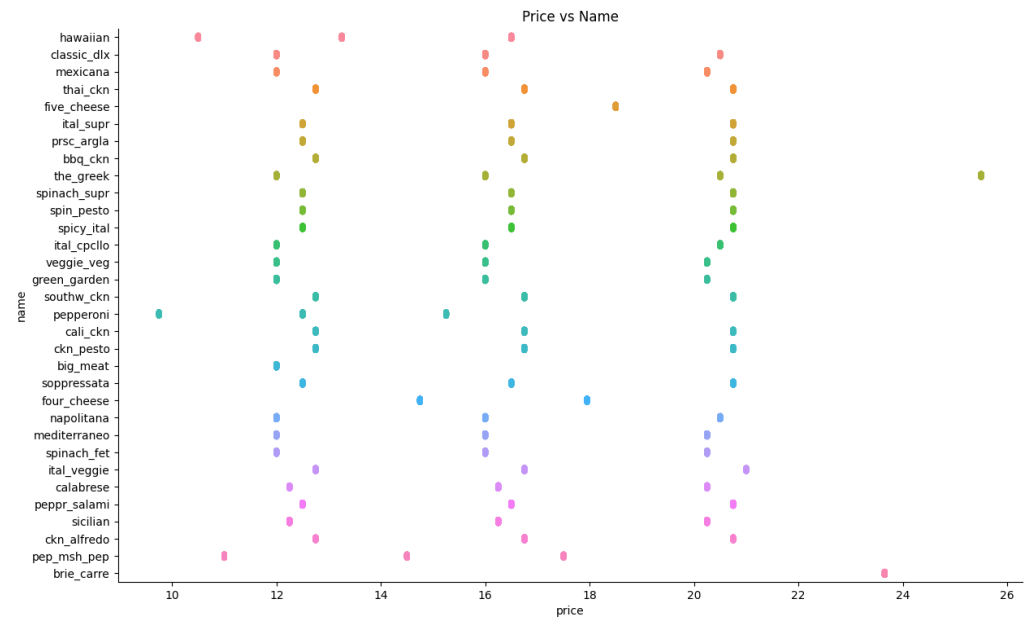
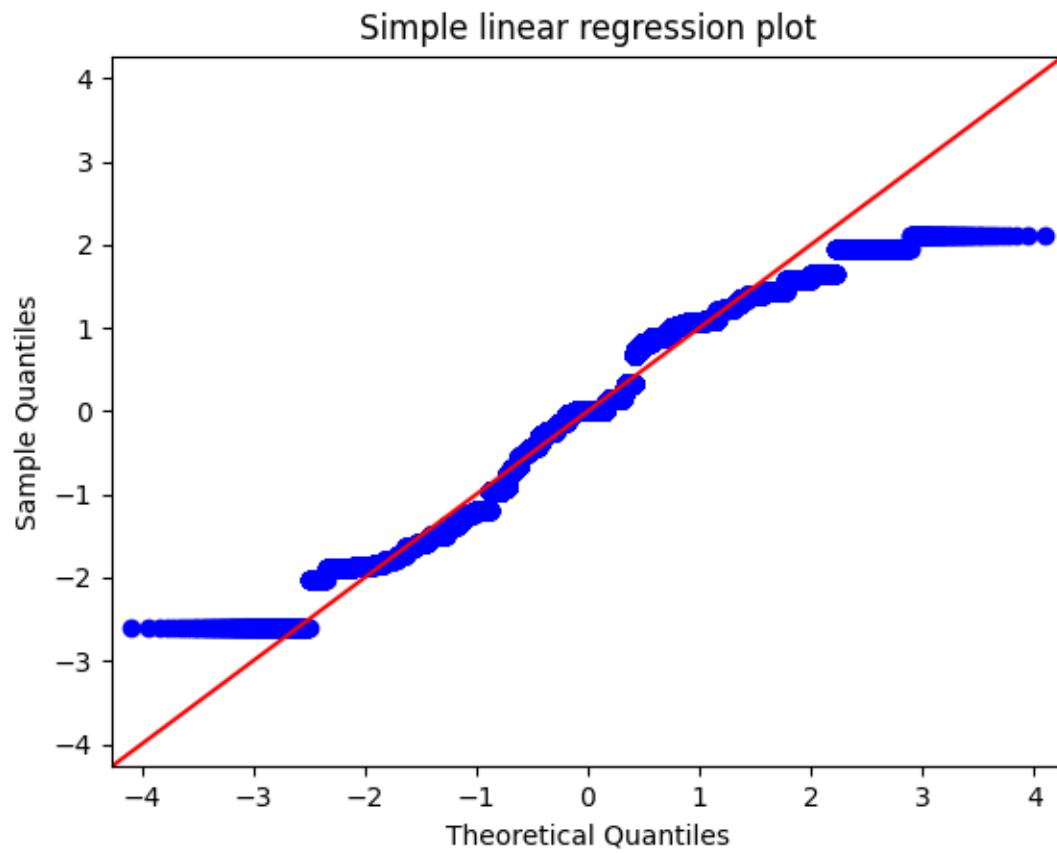
Omnibus:	4627.087	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1698.323
Skew:	-0.211	Prob(JB):	0.00
Kurtosis:	2.197	Cond. No.	26.7

=====

=====

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



Cross table of price of pizza vs name of pizza purchased:

name bbq_ckn big_meat brie_carre ... thai_ckn the_greek veggie_veg

price			...			
9.75	0	0	0 ...	0	0	0
10.50	0	0	0 ...	0	0	0
11.00	0	0	0 ...	0	0	0
12.00	0	1914	0 ...	0	304	464
12.25	0	0	0 ...	0	0	0
12.50	0	0	0 ...	0	0	0
12.75	484	0	0 ...	480	0	0
13.25	0	0	0 ...	0	0	0
14.50	0	0	0 ...	0	0	0
14.75	0	0	0 ...	0	0	0
15.25	0	0	0 ...	0	0	0
16.00	0	0	0 ...	0	281	635
16.25	0	0	0 ...	0	0	0
16.50	0	0	0 ...	0	0	0
16.75	956	0	0 ...	481	0	0
17.50	0	0	0 ...	0	0	0
17.95	0	0	0 ...	0	0	0
18.50	0	0	0 ...	0	0	0
20.25	0	0	0 ...	0	0	427
20.50	0	0	0 ...	0	255	0
20.75	992	0	0 ...	1410	0	0
21.00	0	0	0 ...	0	0	0
23.65	0	0	490 ...	0	0	0
25.50	0	0	0 ...	0	552	0

[24 rows x 32 columns]

Display count of pizza name:

	time	size	type	price
name				
bbq_ckn	2432	2432	2432	2432
big_meat	1914	1914	1914	1914
brie_carre	490	490	490	490
calabrese	937	937	937	937
cali_ckn	2370	2370	2370	2370
ckn_alfredo	987	987	987	987
ckn_pesto	973	973	973	973
classic_dlx	2453	2453	2453	2453
five_cheese	1409	1409	1409	1409
four_cheese	1902	1902	1902	1902
green_garden	997	997	997	997
hawaiian	2422	2422	2422	2422
ital_cpello	1438	1438	1438	1438
ital_supr	1884	1884	1884	1884
ital_veggie	981	981	981	981
mediterraneo	934	934	934	934
mexicana	1484	1484	1484	1484
napolitana	1464	1464	1464	1464
pep_msh_pep	1359	1359	1359	1359
pepperoni	2418	2418	2418	2418
peppr_salami	1446	1446	1446	1446
prsc_argla	1457	1457	1457	1457
sicilian	1938	1938	1938	1938
soppressata	961	961	961	961
southw_ckn	1917	1917	1917	1917
spicy_ital	1924	1924	1924	1924

spin_pesto	970	970	970	970
spinach_fet	1446	1446	1446	1446
spinach_supr	950	950	950	950
thai_ckn	2371	2371	2371	2371
the_greek	1392	1392	1392	1392
veggie_veg	1526	1526	1526	1526

Display count of pizza price:

	time	name	size	type
price				
9.75	751	751	751	751
10.50	1020	1020	1020	1020
11.00	578	578	578	578
12.00	5744	5744	5744	5744
12.25	850	850	850	850
12.50	3380	3380	3380	3380
12.75	2529	2529	2529	2529
13.25	483	483	483	483
14.50	397	397	397	397
14.75	586	586	586	586
15.25	728	728	728	728
16.00	4522	4522	4522	4522
16.25	1136	1136	1136	1136
16.50	4111	4111	4111	4111
16.75	4380	4380	4380	4380
17.50	384	384	384	384
17.95	1316	1316	1316	1316
18.50	1409	1409	1409	1409
20.25	3093	3093	3093	3093

20.50 2026 2026 2026 2026

20.75 8891 8891 8891 8891

21.00 190 190 190 190

23.65 490 490 490 490

25.50 552 552 552 552