IDEALISTA

Data Analytics

By Jacky Barraza

About Jacky

Data – Idealista
Quality

Data Analytics
Price Houses in Madrid

Architecture (cloud) I
Azure

Architecture II
Azure

Hazards and Mitigation
External web source

HOUSE PRICE IN MADRID

2019 - 2020

idealista

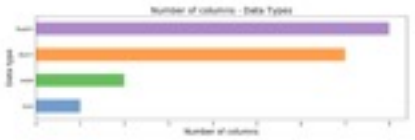
idealista is in the market since 2000
House listing website for buying and renting

Dataset, Madrid houses offered in idealista.com late 2019 - 2020:

DATASET

18 Columns
41,961 Rows

NUMERIC VAR	CATEGORICAL VAR
56 % 11 var	43 % 7 var



HOUSE PRICE DATASET QUALITY DATA

Histograms and boxplots to have a general look to sample distribution in a variable

CATEGORICAL VARIABLES

Source URL – 41,961	Elevator_class – 2
Quarter – 136	Ext_Int – 2
District – 21	New – 2
Garage_opt_inc – 3	

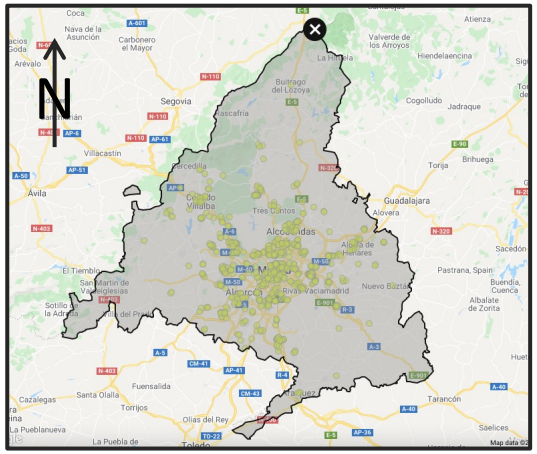
FINDINGS

Outliers
Null Values - New
High number of Classes in the categorical variables
No duplicates

HOUSE PRICE IN MADRID

2019 - 2020

idealista



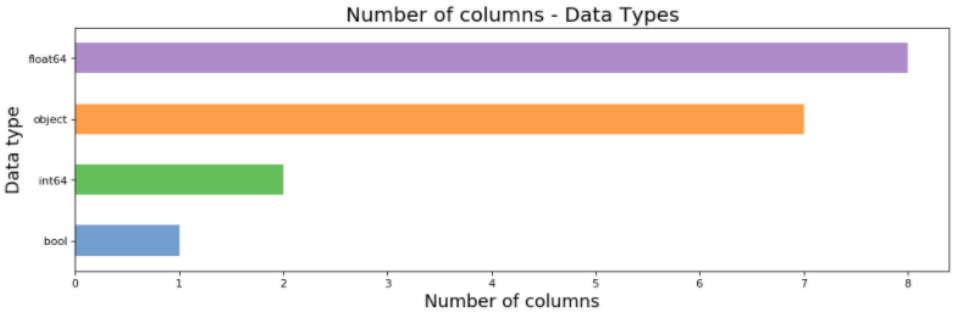
idealista is in the market since 2000
House listing website for buying and renting

Dataset, Madrid houses offered in idealista.com late 2019 - 2020:

DATASET

18 Columns
41,961 Rows

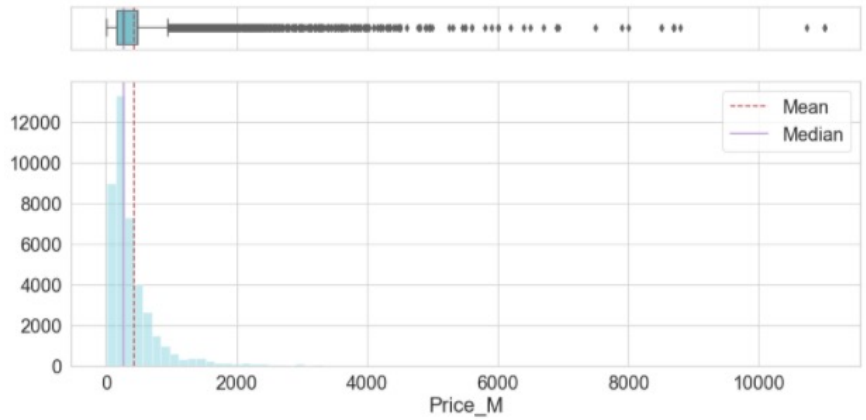
NUMERIC VAR	CATEGORICAL VAR
56 % 11 var	43 % 7 var



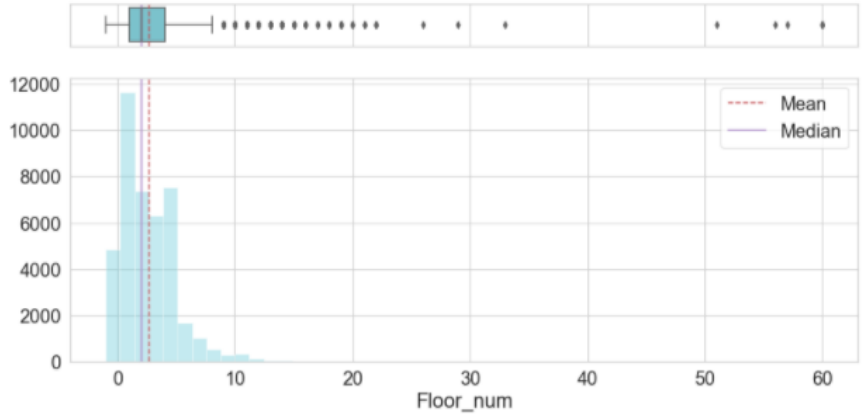
HOUSE PRICE DATASET QUALITY DATA

Histograms and boxplots to have a general look to sample distribution in a variable

Price/1000: Price_M



Floor_num



CATEGORICAL VARIABLES

Source URL – 41,961
Quarter – 136
District – 21
Garage_opt_inc – 3
Elevator_class – 2
Ext_int – 2
New – 2

FINDINGS

Outliers
Null Values - New
High number of Classes in the categorical variables
No duplicates


Banesto
BBVA


MEAN AND MEDIAN
MADRID PRICE HOUSE
Price_M Statistics

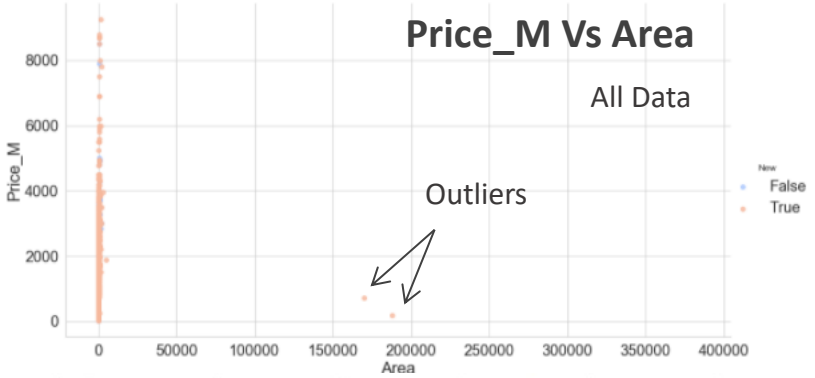

PRICE CORRELATIONS
MADRID PRICE HOUSE

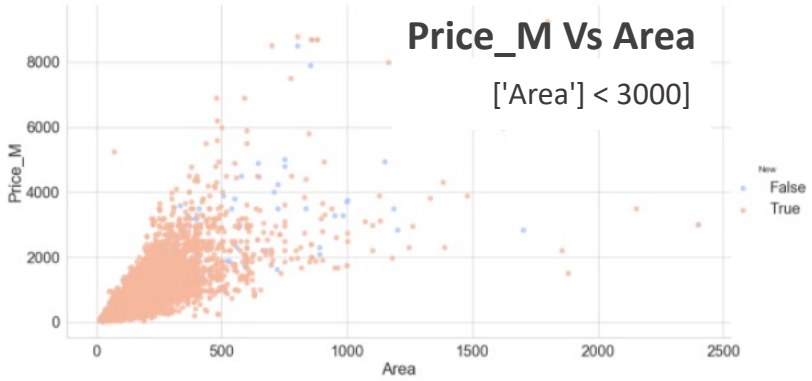

FLOORS AND AREA (square mts)
MADRID PRICE HOUSE


PRICE PER SQUARE MTS EVOLVED 2020
MADRID PRICE HOUSE

MEAN AND MEDIAN

MADRID PRICE HOUSE





Price_M Statistics

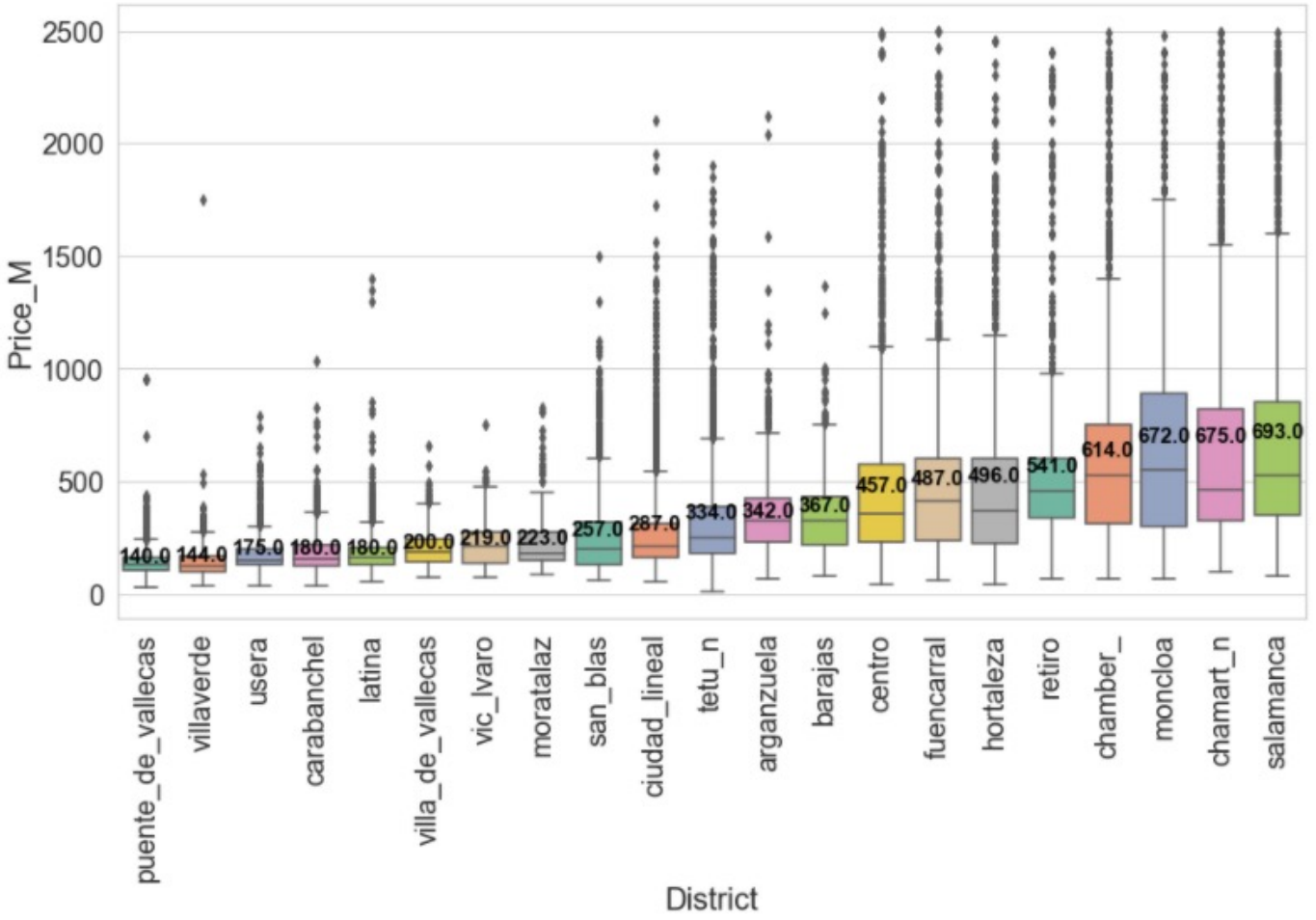
LIMITS	MEAN	MEDIAN
Price: 13K – 11,007K Area: 12 – 385,000 Samples: 41,961	426.01 std: 515.34	270.0

LIMITS	MEAN	MEDIAN
Price: 13K – 11,007K Area: 12 – 2,400 Samples: 41,956 (-5)	425.88 std: 515.34	270.0

LIMITS	MEAN	MEDIAN
Price: 13K – 2,495K Area: 12 – 1387 Samples: 41,943 (-18)	389.28 std: 362.66	270.0

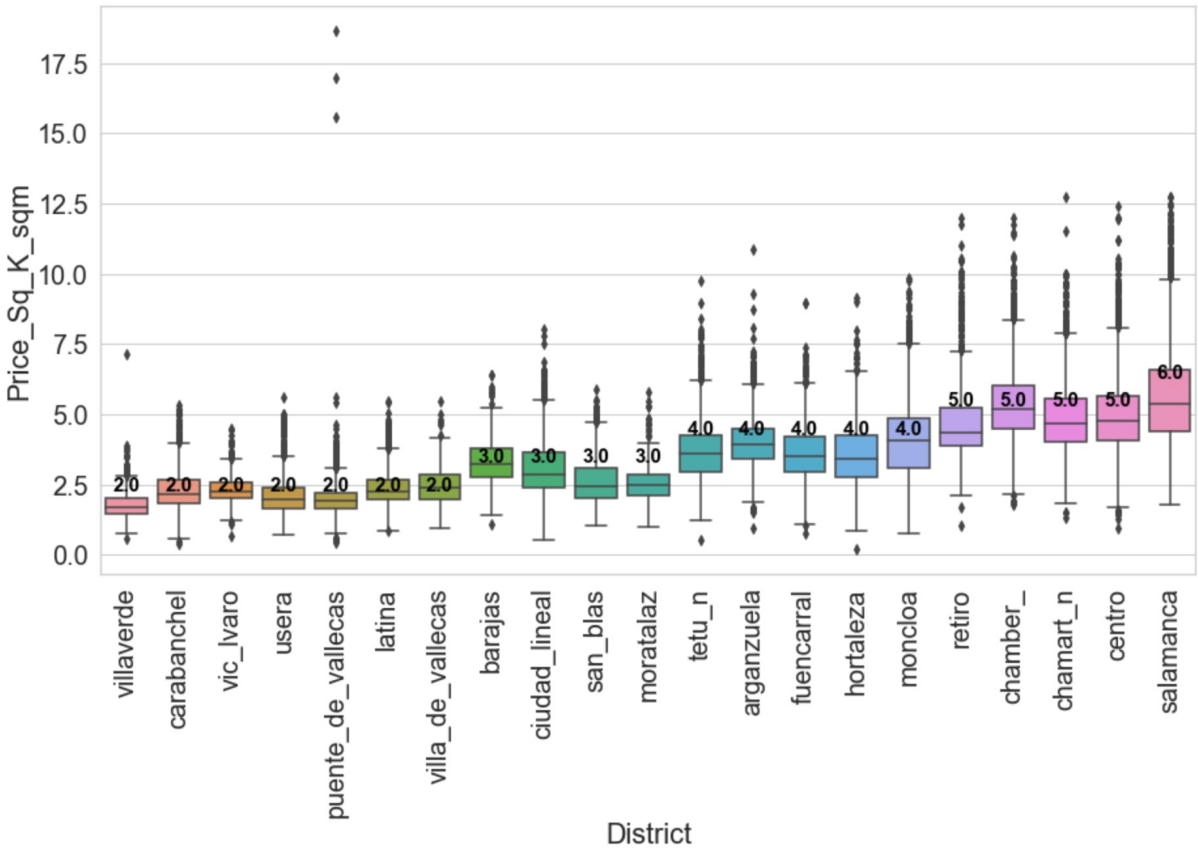
MEAN AND MEDIAN
MADRID PRICE HOUSE

- ✓ The lowest mean is **140K** in Puente de Vallecas
- ✓ The highest mean is **693K** in Salamanca
- ✓ The closest to the mean of all data **389K** are:
  - Arganzuela **342K**
  - Barajas **367K**
  - Tete_n **334K**



MEAN AND MEDIAN
MADRID PRICE HOUSE

- ✓ The lowest mean is **2K** in Villaverde
- ✓ The highest mean is **6K** in Salamanca

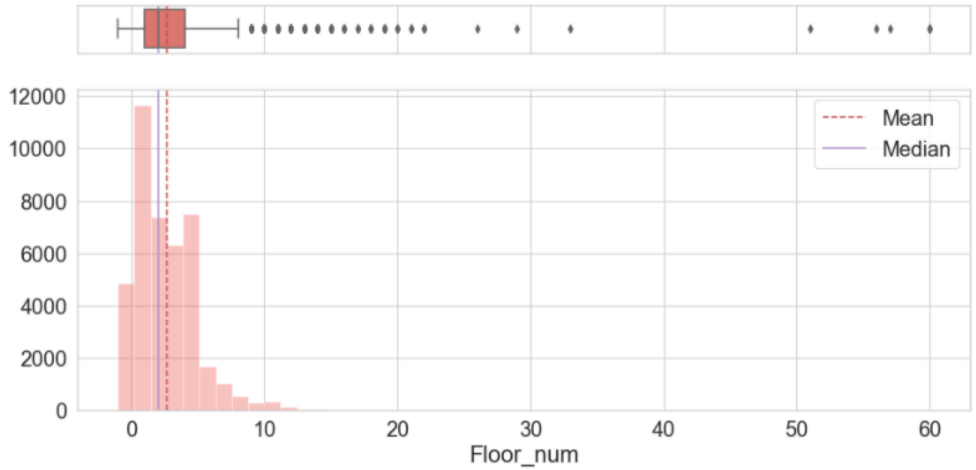


FLOORS AND AREA (square mts)

MADRID PRICE HOUSE

The Histograms show a Skewed distribution

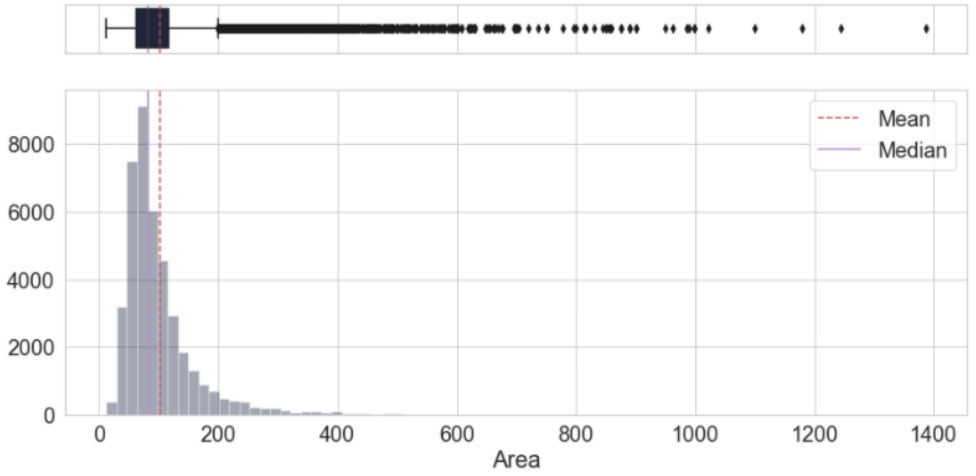
Floor_Num Distribution



LIMITS
Floor_room : (-1) – 60

MEAN	MEDIAN
2.63 std: 2.39	2

Area Distribution



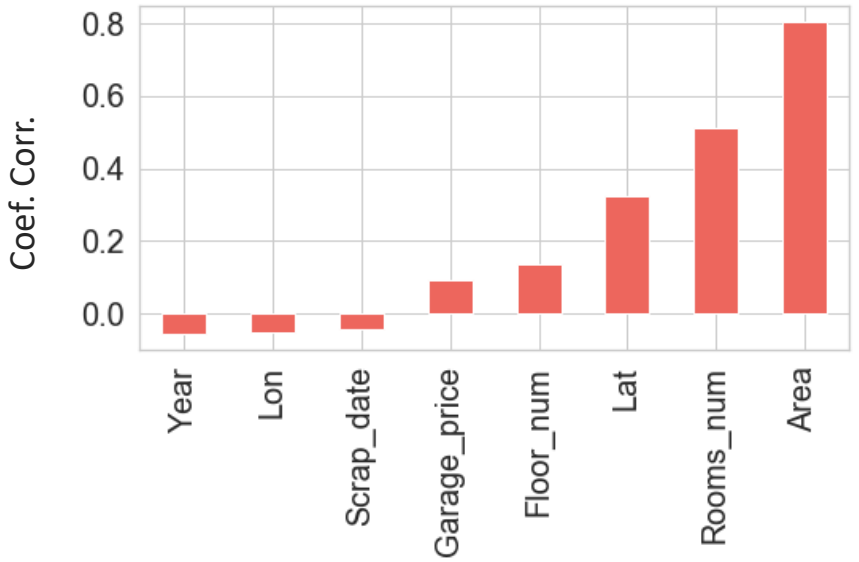
LIMITS
Area: 12 – 1,387

MEAN	MEDIAN
102.76 std: 73.45	82

PRICE CORRELATIONS

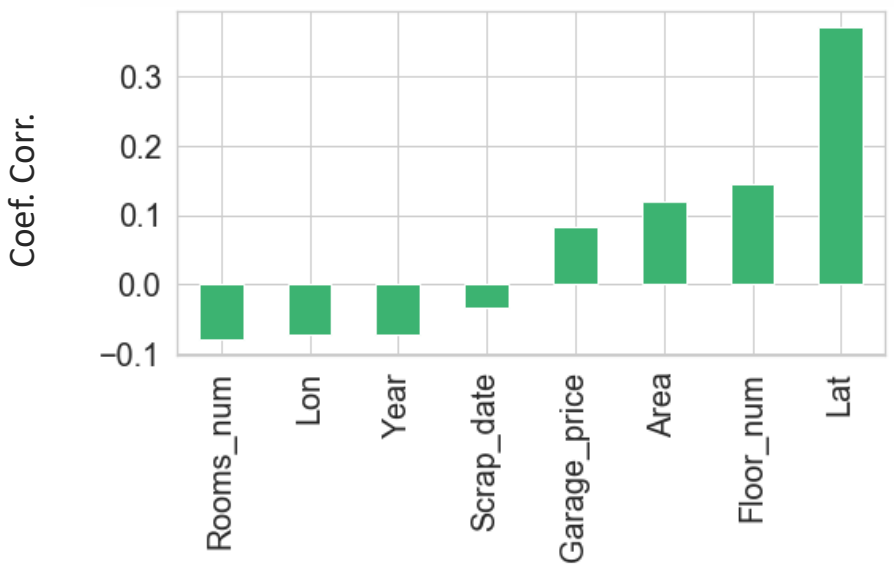
MADRID PRICE HOUSE

Variable correlations against Price_M



Numeric Variable - Price

Variable correlations against Price_Sqmts

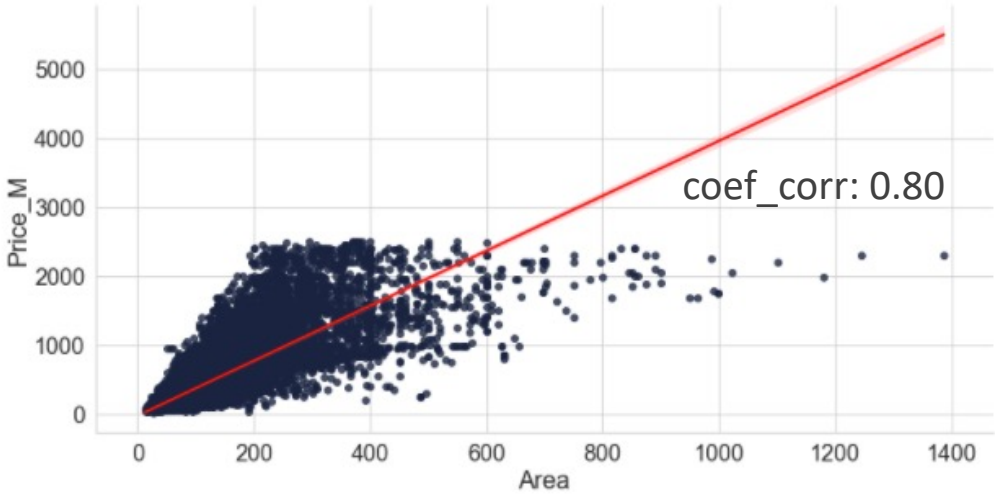


Numeric Variable – Price_Sqmts

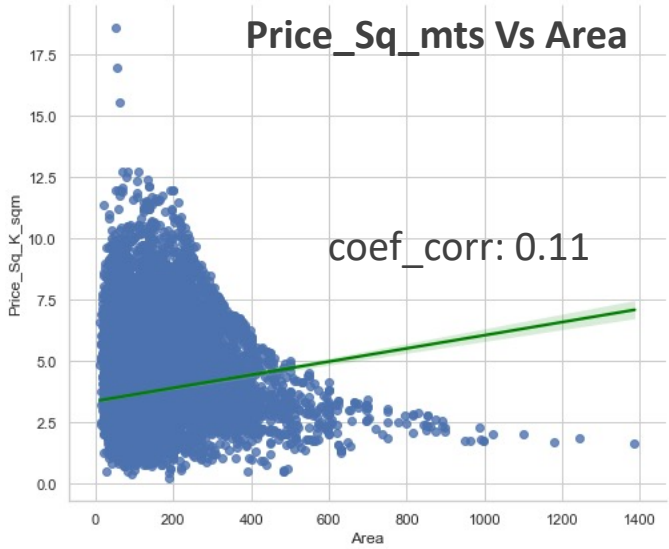
PRICE CORRELATIONS

MADRID PRICE HOUSE

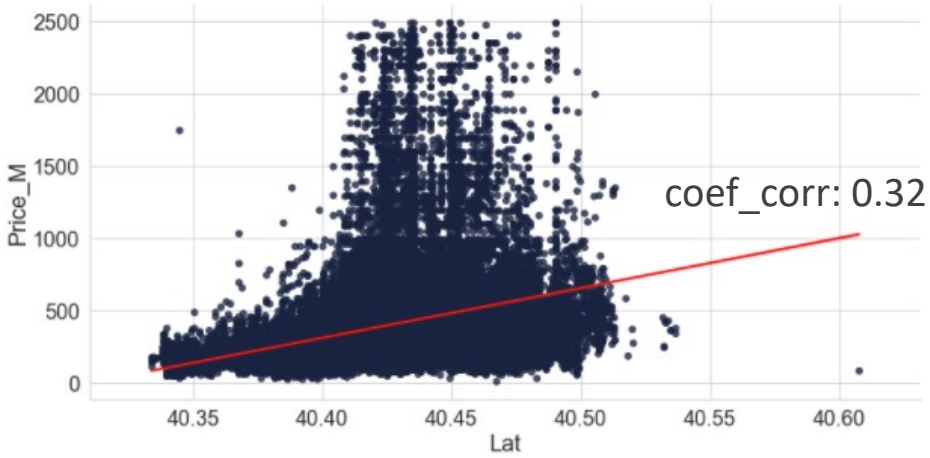
Price_M Vs Area



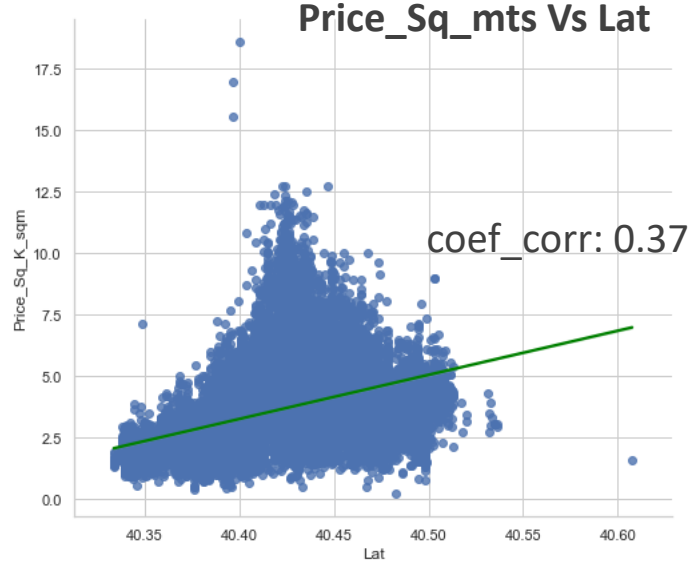
Price_Sq_mts Vs Area



Price_M Vs Lat



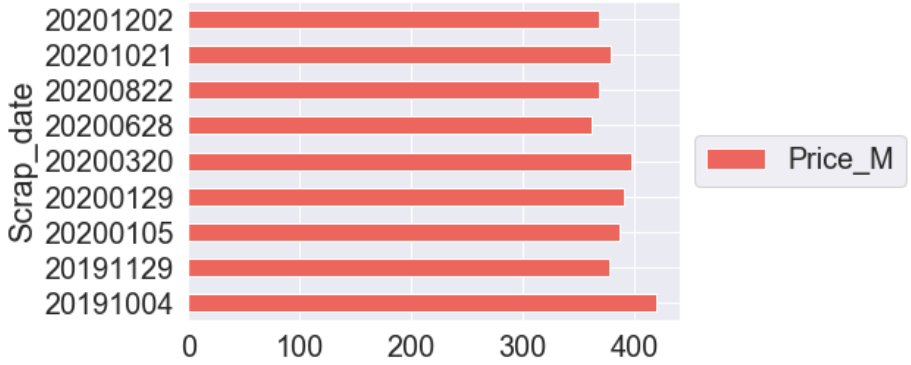
Price_Sq_mts Vs Lat

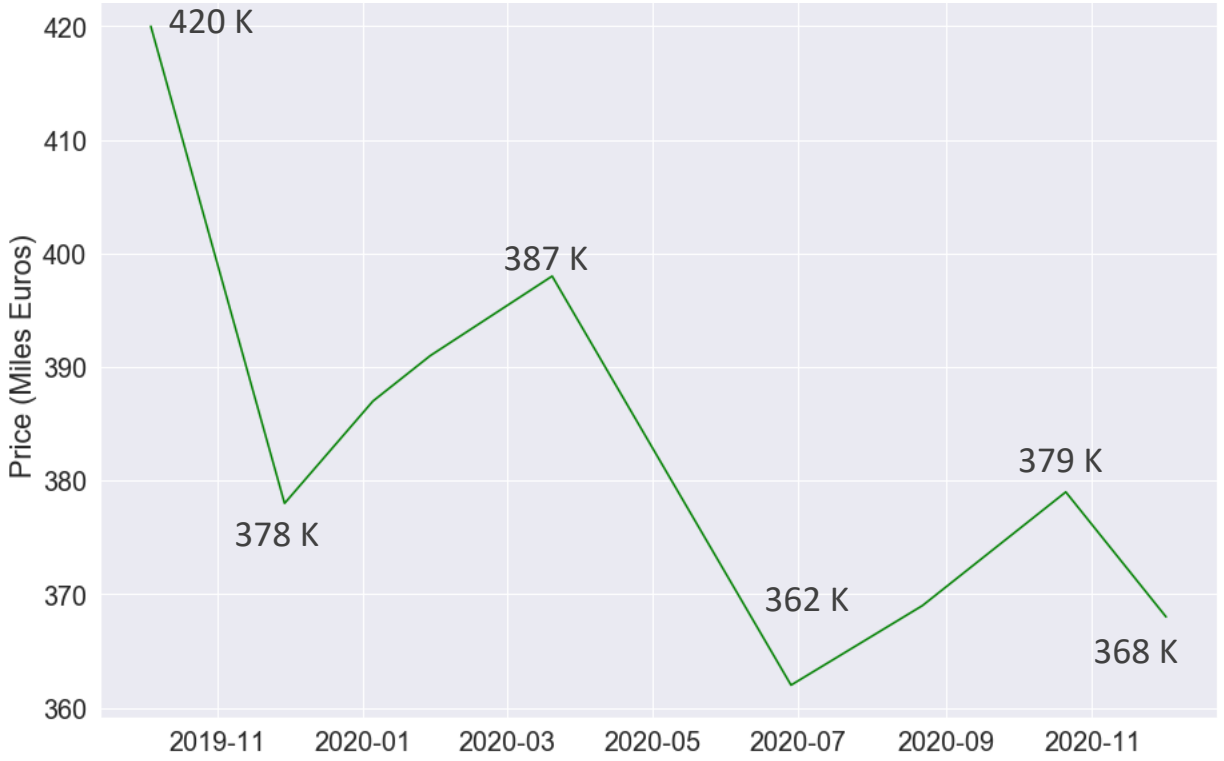


PRICE PER SQUARE MTS EVOLVED 2020

MADRID PRICE HOUSE

Price_M aggrupation by Scrap_date

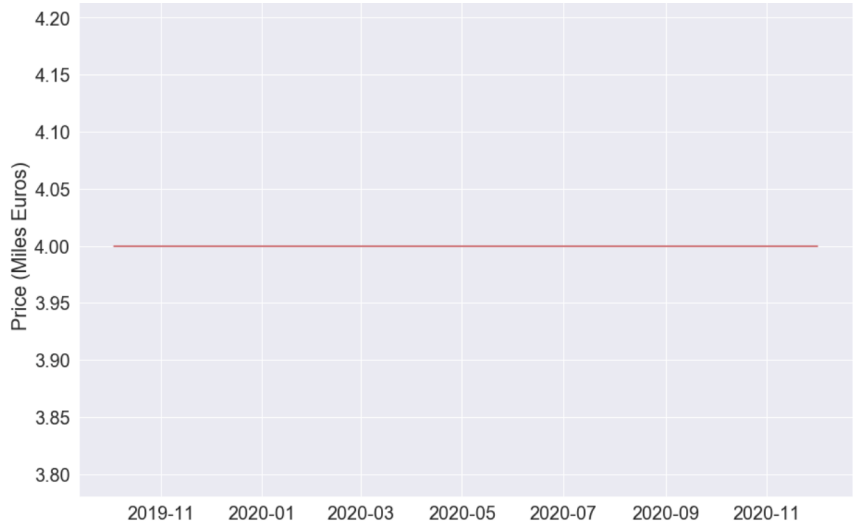




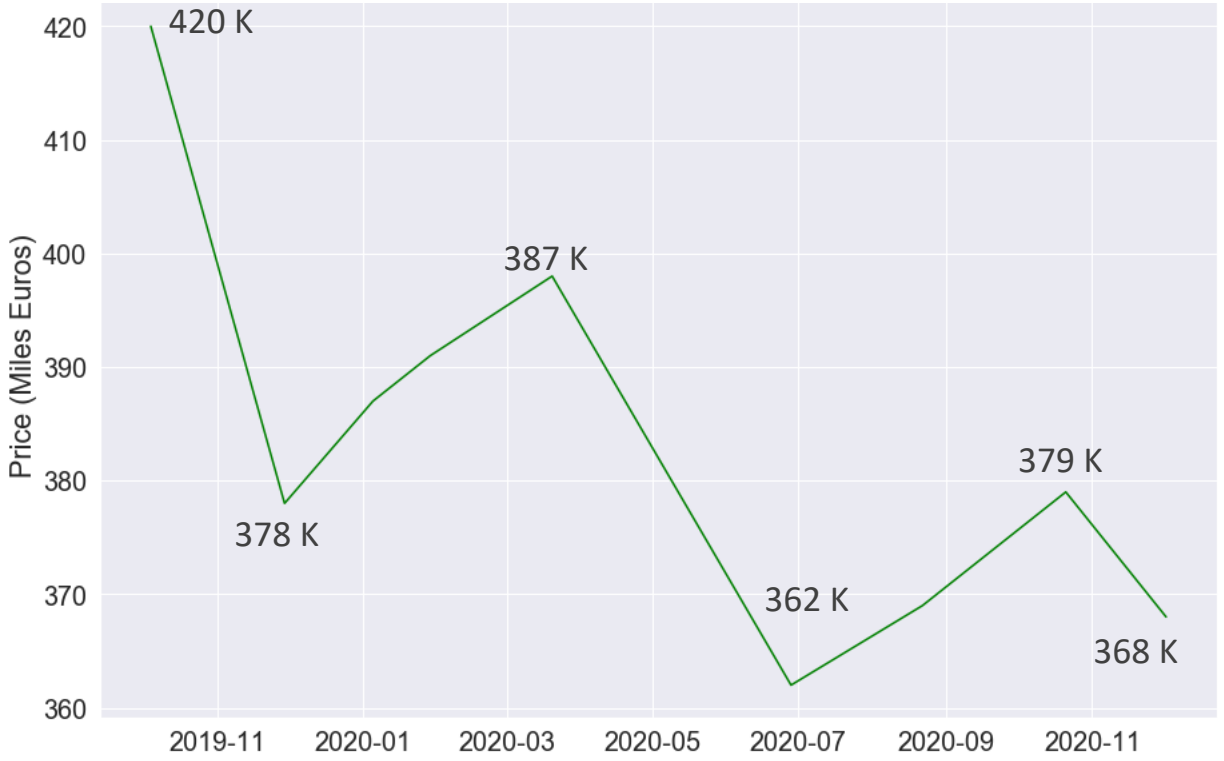
PRICE PER SQUARE MTS EVOLVED 2020

MADRID PRICE HOUSE

Price_M/Sqmts aggrupation by Scrap_date



Price_M aggrupation by Scrap_date

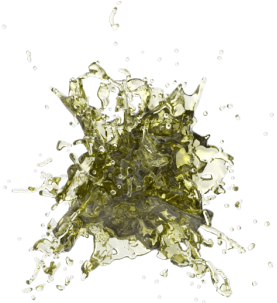


CLOUD ARQUITECURE PROPOSAL

MADRID PRICE HOUSE

NEEDS:

- ❑ Allow repeatable data collection (scraping)
- ❑ Storing data (might needs images, videos…)
- ❑ Connection with data analytics, data scientist (real time – long term)
- ❑ Hazards mitigation (web sources)

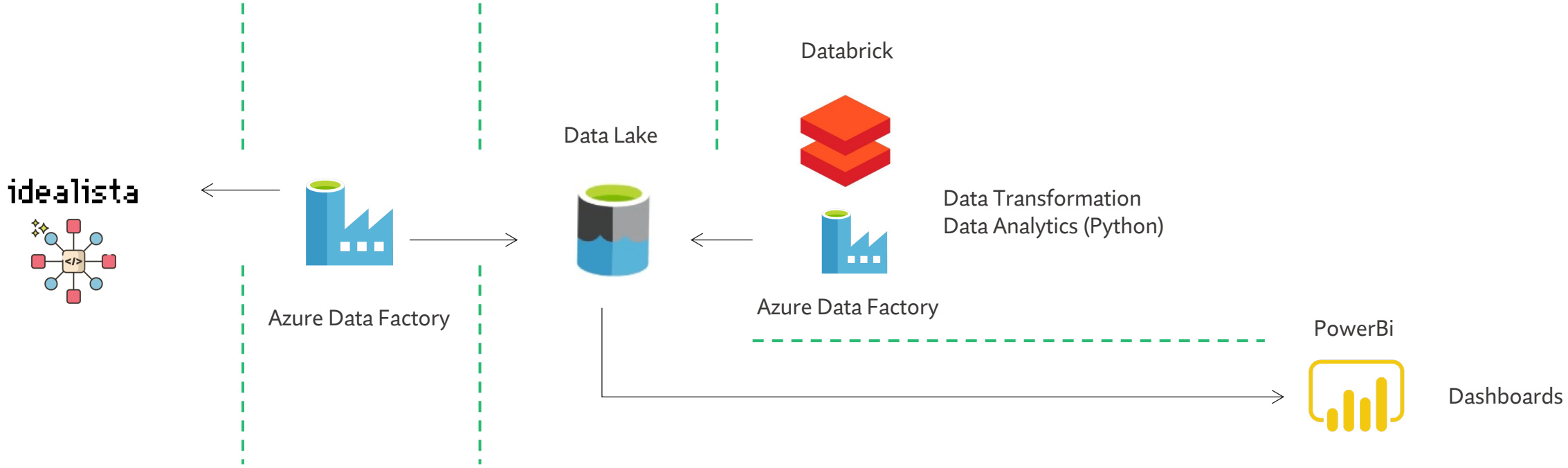


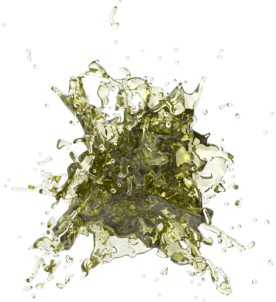
CLOUD ARQUITECURE I

PROPOSAL

MADRID PRICE HOUSE

Data Source	Loading	Landing	Analysis and Transformation	Data Analytics



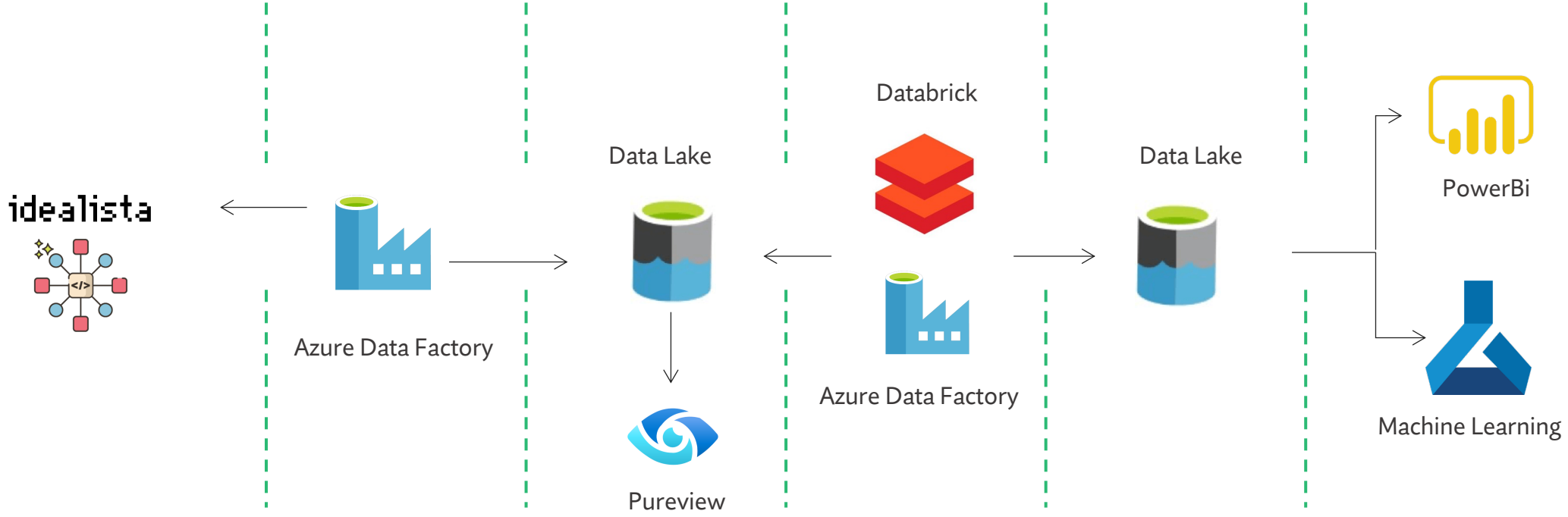


CLOUD ARQUITECURE II

PROPOSAL

MADRID PRICE HOUSE

Data Source	Loading	Landing	Analysis and Transformation	Datawarehouse	Potential Use Cases



SET UP AN ENVIRONMENT TO ALLOW REPEATABLE DATA COLLECTION (SCRAPING)

MADRID PRICE HOUSE

Data Source | **Loading** | **Landing**

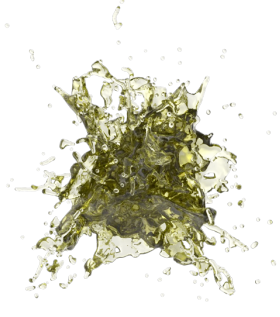
idealista

Azure Data Factory

Data Lake

Azure Data Factory

Build pipelines to trigger data collection every month and storage raw data into the Datalake.



DATA STORAGE RECOMENDATIONS

MADRID PRICE HOUSE

Consideration: videos, images, AutoCAD files…

Blob Storage

- ✓ Support Streaming and random access scenarios
- ✓ Object store with flat namespace
- ✓ Access app data from anywhere
- ✓ Datalake – big data analytics

Data Lake

- ✓ Optimized storage for big data analytics workloads
- ✓ Hierarchical file system
- ✓ Unexpected failures - Data lake spreads parts of a file over a number of individual storage servers

Cosmos DB

- ✓ Document store, graph, key-value store, wide column store
- ✓ Guarantees single-digit-millisecond latencies at the 99th percentile anywhere in the world
- ✓ Guarantees high availability with multi-homing
- ✓ No images – No Videos

DATA AVAILABILITY (real-time and long term)

MADRID PRICE HOUSE

AZURE – Service-level agreement (SLA)

Monthly Availability Percentage

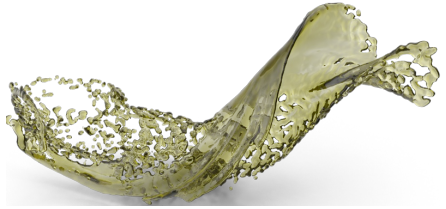
Microsoft Azure

MONTHLY READ AVAILABILITY PERCENTAGE	SERVICE CREDIT
< 99.999%	10%
< 99%	25%

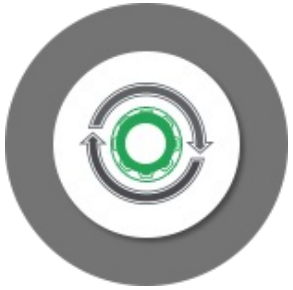
EXTERNAL WEB SOURCES

Possible issues and mitigation

MADRID PRICE HOUSE

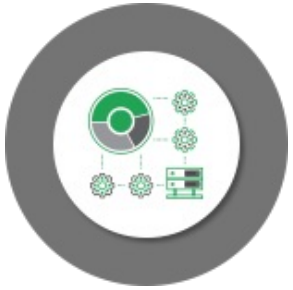


Dynamic content

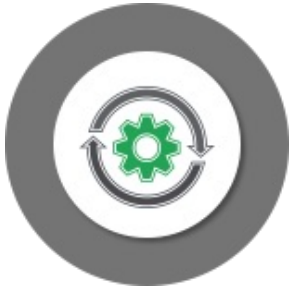


Complicated and changeable web pages structure

Real-Time



Slow/unstable load speed

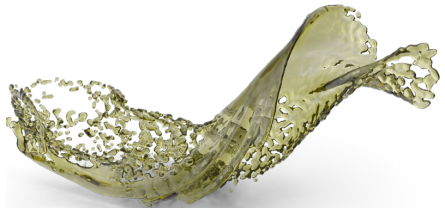


- ✓ Create a system when data strucutre is diferent (COSMO DB handle this issues ok)
- ✓ Azure Event Hub (streaming data)
  - Anomaly detection (fraud/outliers)
  - Application logging
  - Analytics pipelines, such as clickstreams
  - Live dashboarding
  - Archiving data
  - Transaction processing
  - User telemetry processing
  - Device telemetry streaming

CLOUD ARQUITECURE BUDGET

MADRID PRICE HOUSE

Microsoft Azure Estimate					
Advanced Analytics on Big Data					
Service type	Custom name	Region	Description	Estimated monthly cost	Estimated upfront cost
Data Factory		East US 2	Azure Data Factory V2 Type, Data Pipeline Service Type, Azure Integration Runtime: 100 Activity Run(s), 100 Data movement unit(s), 10,000 Pipeline activities, 10,000 Pipeline activities – External, Self-hosted Integration Runtime: 100 Activity Run(s), 1,000 Data movement unit(s), 10,000 Pipeline activities, 10,000 Pipeline activities – External, Data Flow: 1 x 8 Compute Optimized vCores x 730 Hours, 1 x 8 General Purpose vCores x 730 Hours, 1 x 8 Memory Optimized vCores x 730 Hours, Data Factory Operations: 100 x 50,000 Read/Write operation(s), 100 x 50,000 Monitoring operation(s)	$5,045.12	$0.00
Power BI Embedded		East US 2	1 node(s) x 720 Hours, Node type: A2, 2 Virtual Core(s), 5GB RAM, 301-600 Peak renders/hour	$1,445.83	$0.00
Azure Databricks		West US	All-Purpose Compute Workload, Premium Tier, 1 D3V2 (4 vCPU(s), 14 GB RAM) x 730 Hours, Pay as you go, 0.75 DBU x 730 Hours	$504.80	$0.00
Azure Data Lake Storage Gen1		East US 2	Pay-as-you-go: 200 GB Storage, 500 Read Transactions, 500 Write Transactions	$34.80	$0.00
Data Factory		East US	Azure Data Factory V2 Type, Data Pipeline Service Type,	$0.00	$0.00
Support			**Support**	$0.00	$0.00
			Licensing Program	**Microsoft Online Services Agreement**	
			Total	$7,030.22	$0.00
Disclaimer					
All prices shown are in US Dollar ($). This is a summary estimate, not a quote. For up to date pricing information please visit https://azure.microsoft.com/pricing/calculator/					
This estimate was created at 3/15/2021 9:05:27 PM UTC.					



GRACIAS

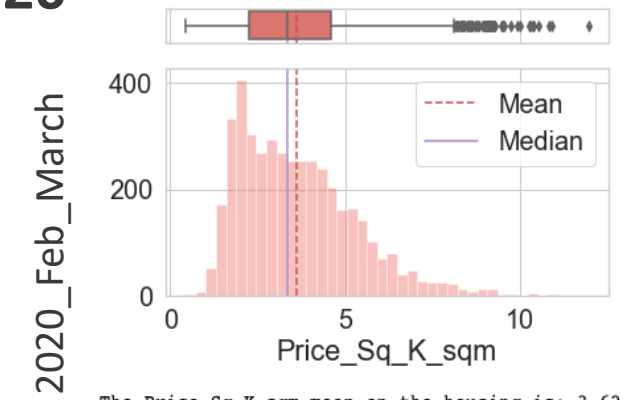
Preguntas

PRICE PER SQUARE MTS EVOLVED 2020

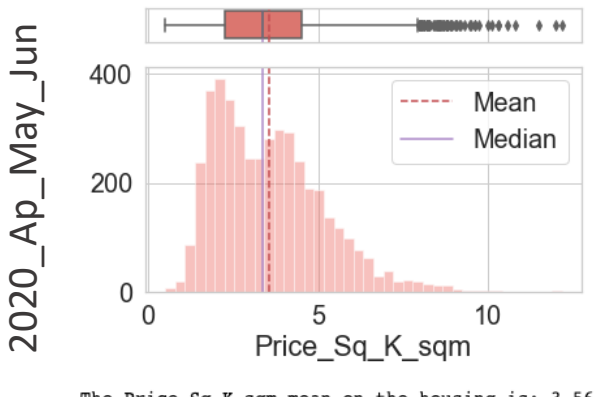
MADRID PRICE HOUSE

Price by Sqmts aggrupation by Scrap_date

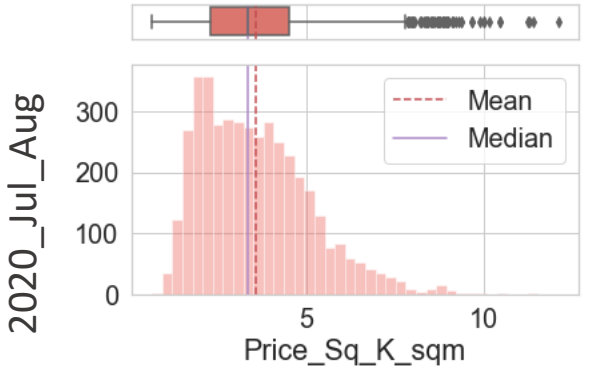
En 1.900 horas con 5 personas en el equipo se alcanzó a desarrollar:



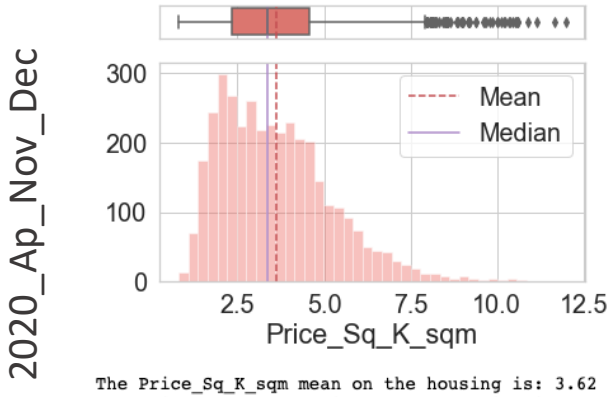

```
The Price_Sq_K_sqm mean on the housing is: 3.62
The Price_Sq_K_sqm median on the housing is: 3.36
The Price_Sq_K_sqm standard deviation on the housing is: 1.66
```



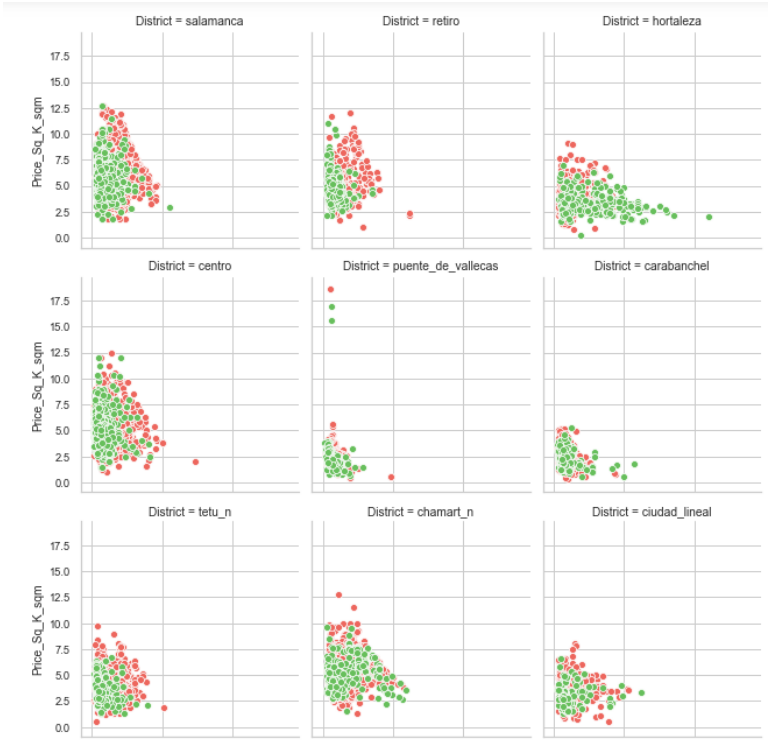

```
The Price_Sq_K_sqm mean on the housing is: 3.56
The Price_Sq_K_sqm median on the housing is: 3.36
The Price_Sq_K_sqm standard deviation on the housing is: 1.61
```

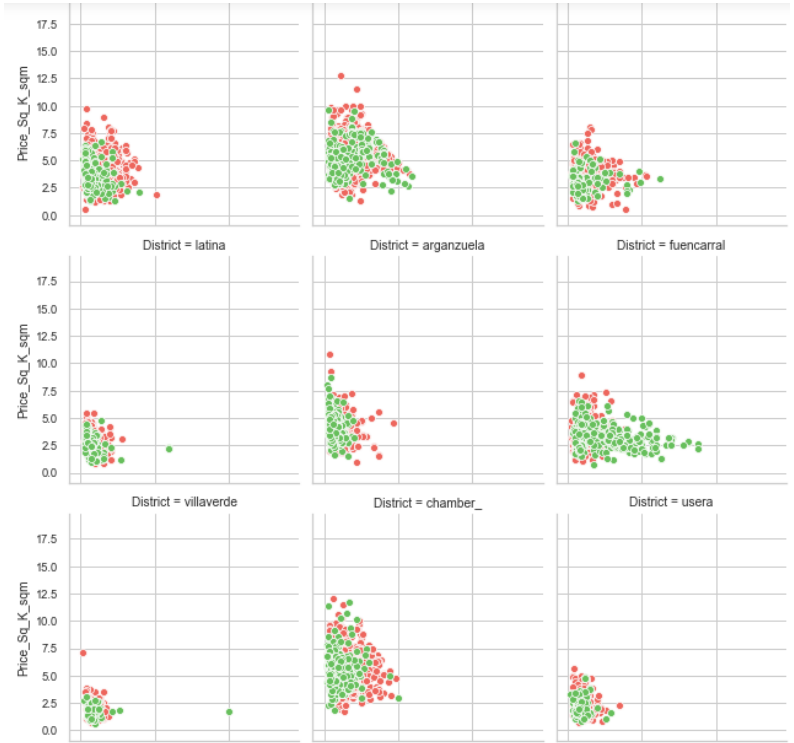


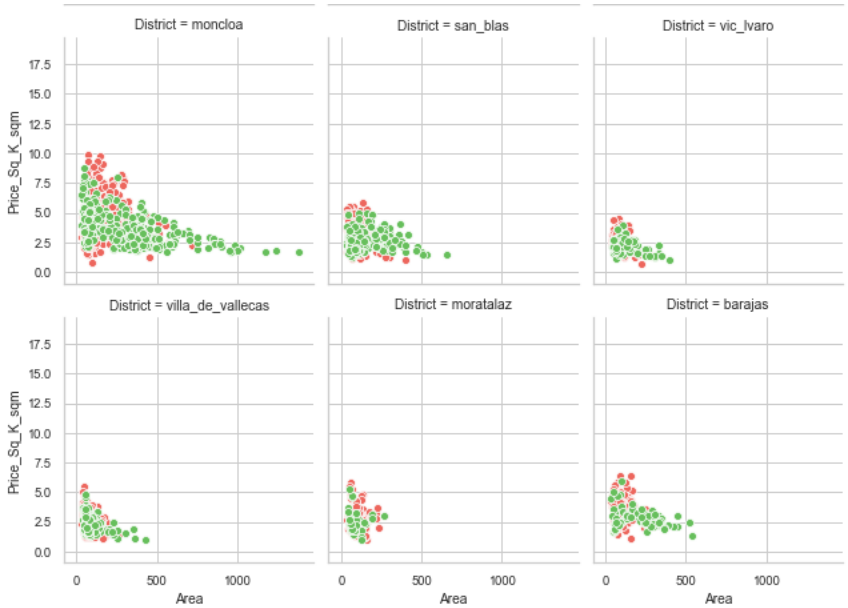

```
The Price_Sq_K_sqm mean on the housing is: 3.57
The Price_Sq_K_sqm median on the housing is: 3.35
The Price_Sq_K_sqm standard deviation on the housing is: 1.58
```

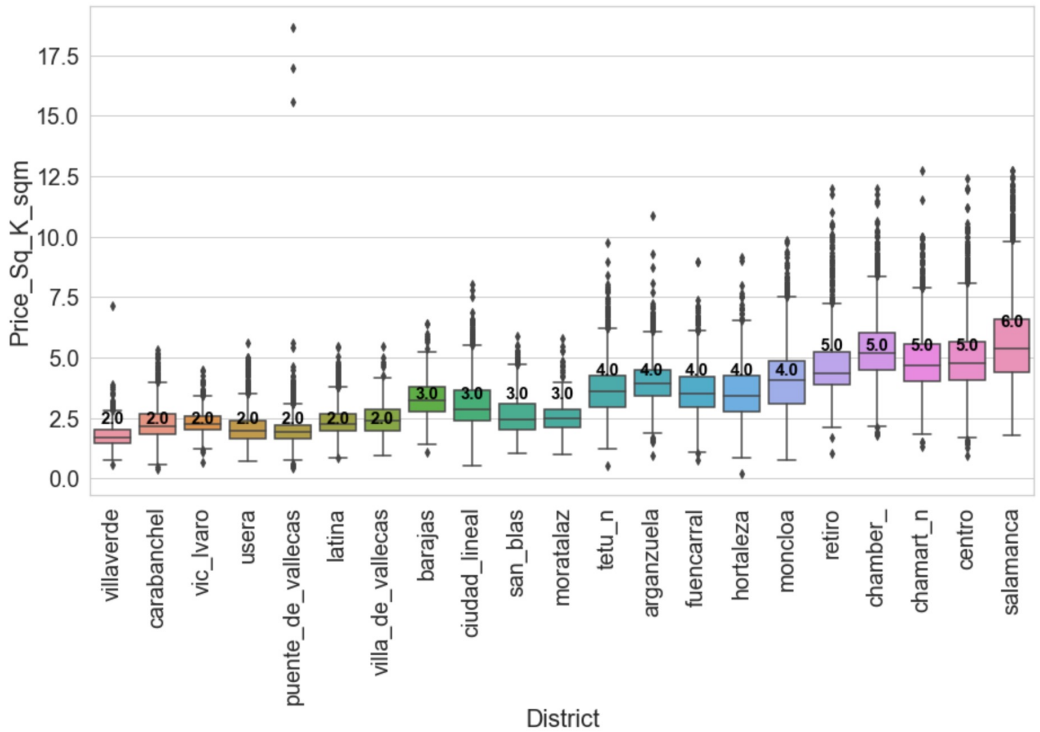


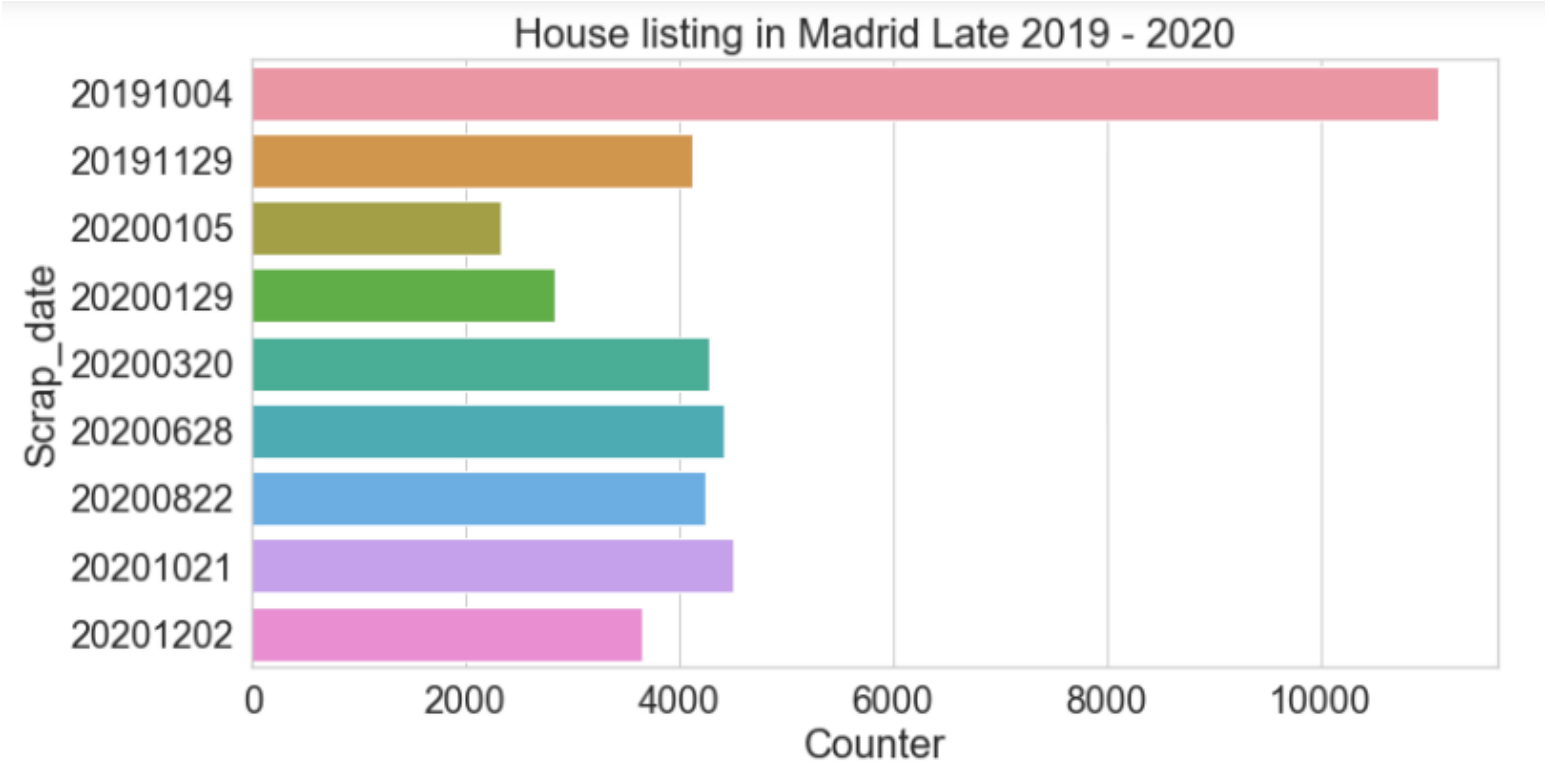

```
The Price_Sq_K_sqm mean on the housing is: 3.62
The Price_Sq_K_sqm median on the housing is: 3.38
The Price_Sq_K_sqm standard deviation on the housing is: 1.64
```

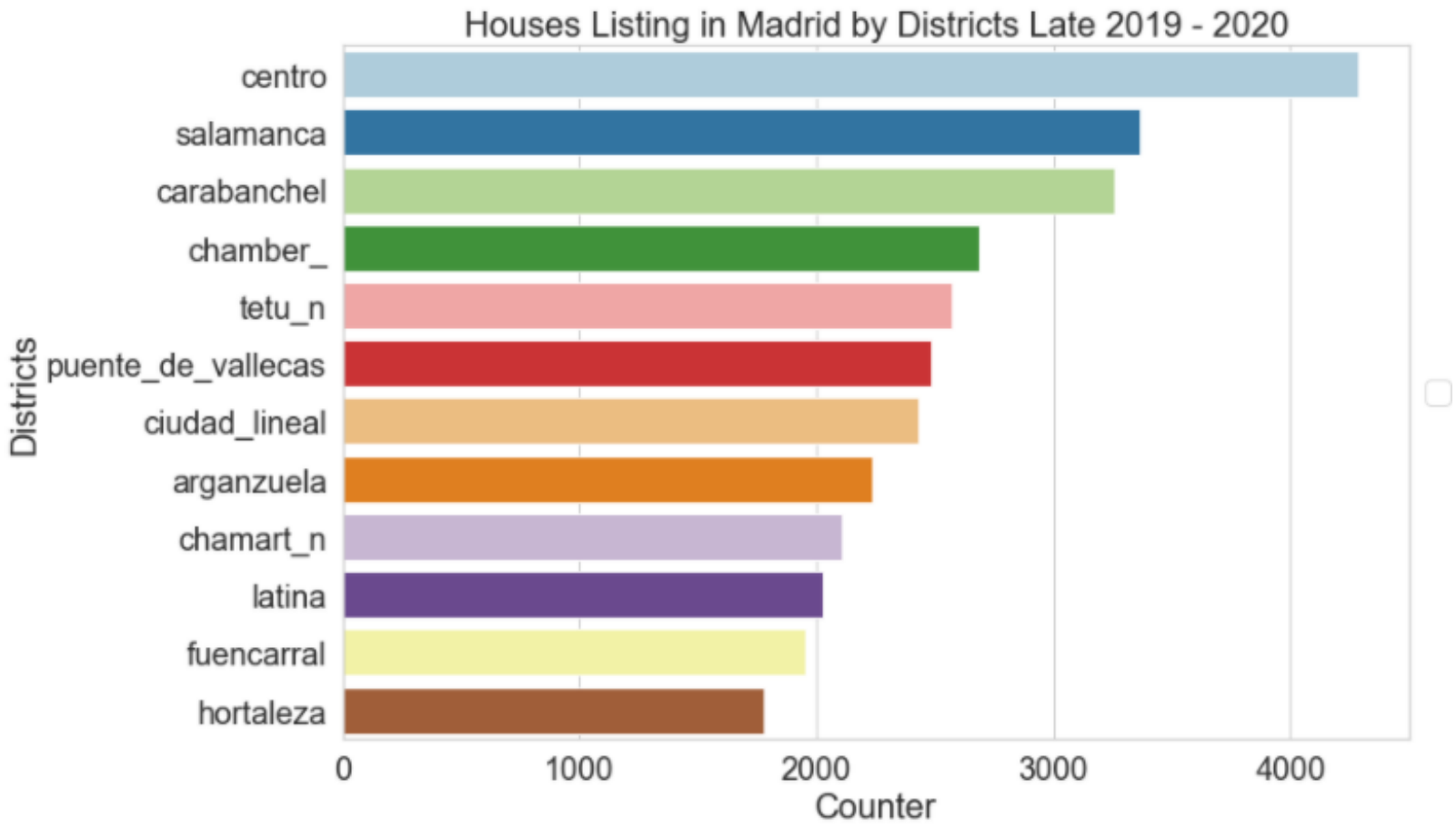

District = salamanca
District = retiro
District = hortaleza
District = centro
District = puente_de_vallecas
District = carabanchel
District = tetu_n
District = chamart_n
District = ciudad_lineal
Price_Sq_K_sqm

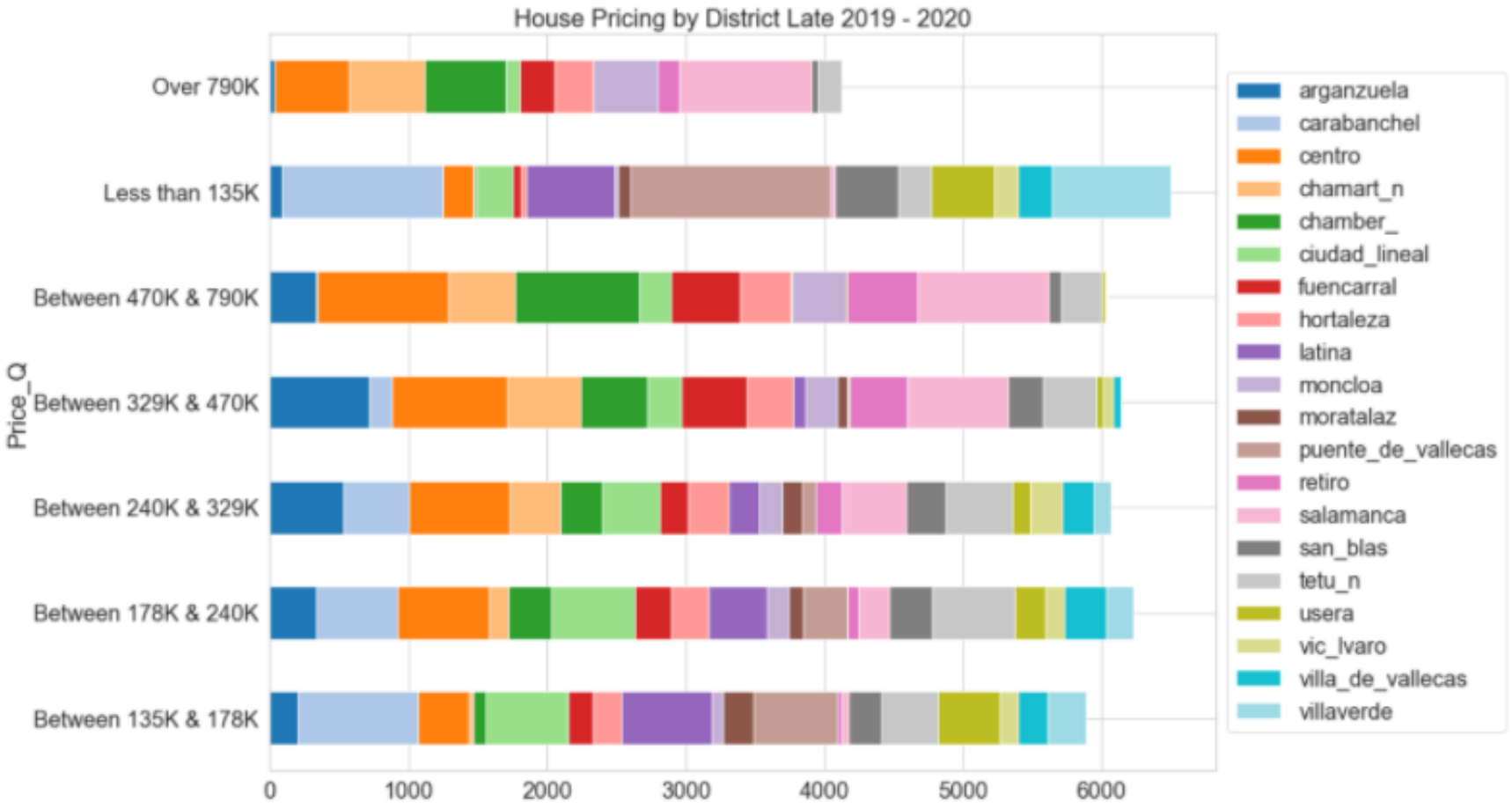

District = latina
District = arganzuela
District = fuencarral
District = villaverde
District = chamber_
District = usera
Price_Sq_K_sqm

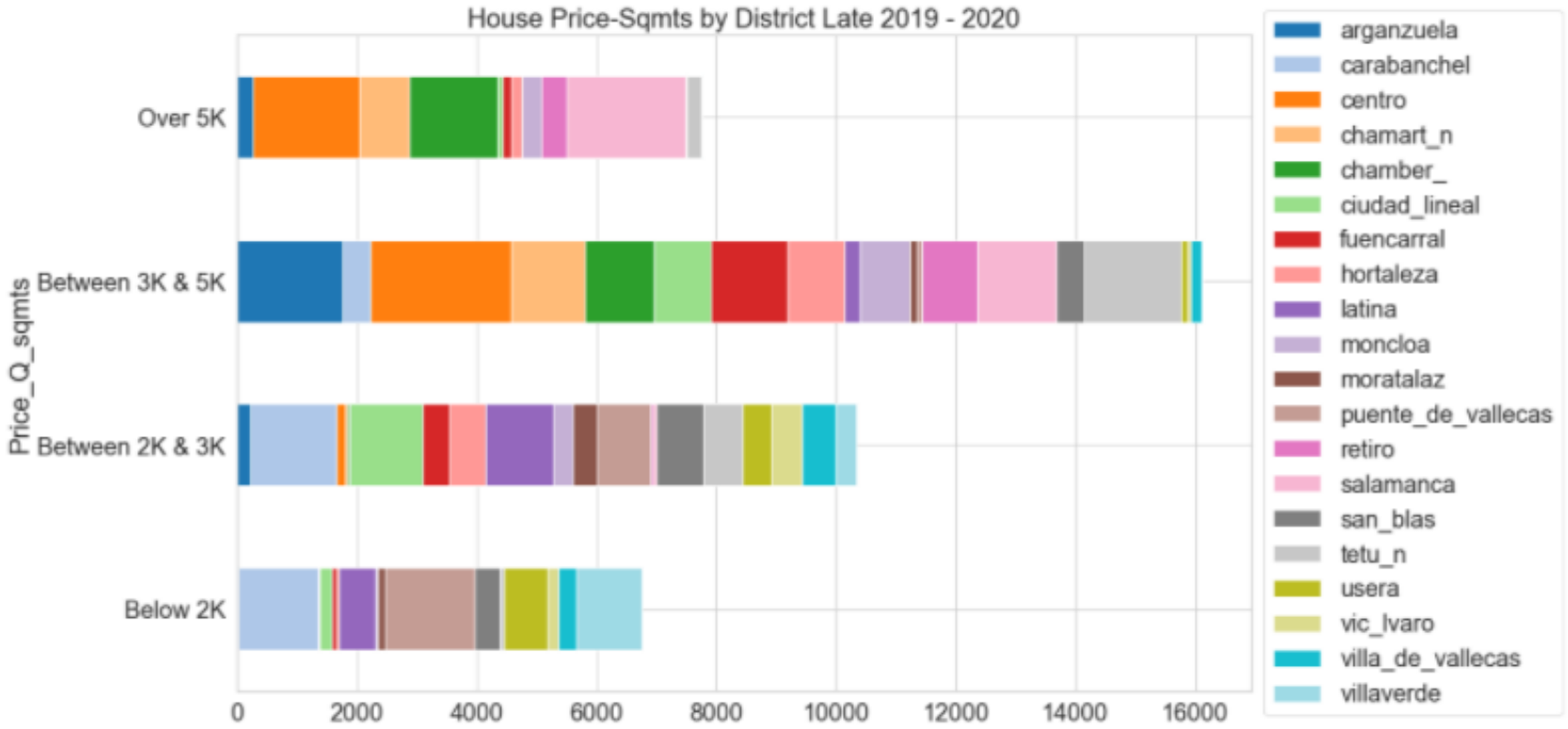

District = moncloa
District = san_blas
District = vic_lvaro
District = villa_de_vallecas
District = moratalaz
District = barajas
Price_Sq_K_sqm
Area

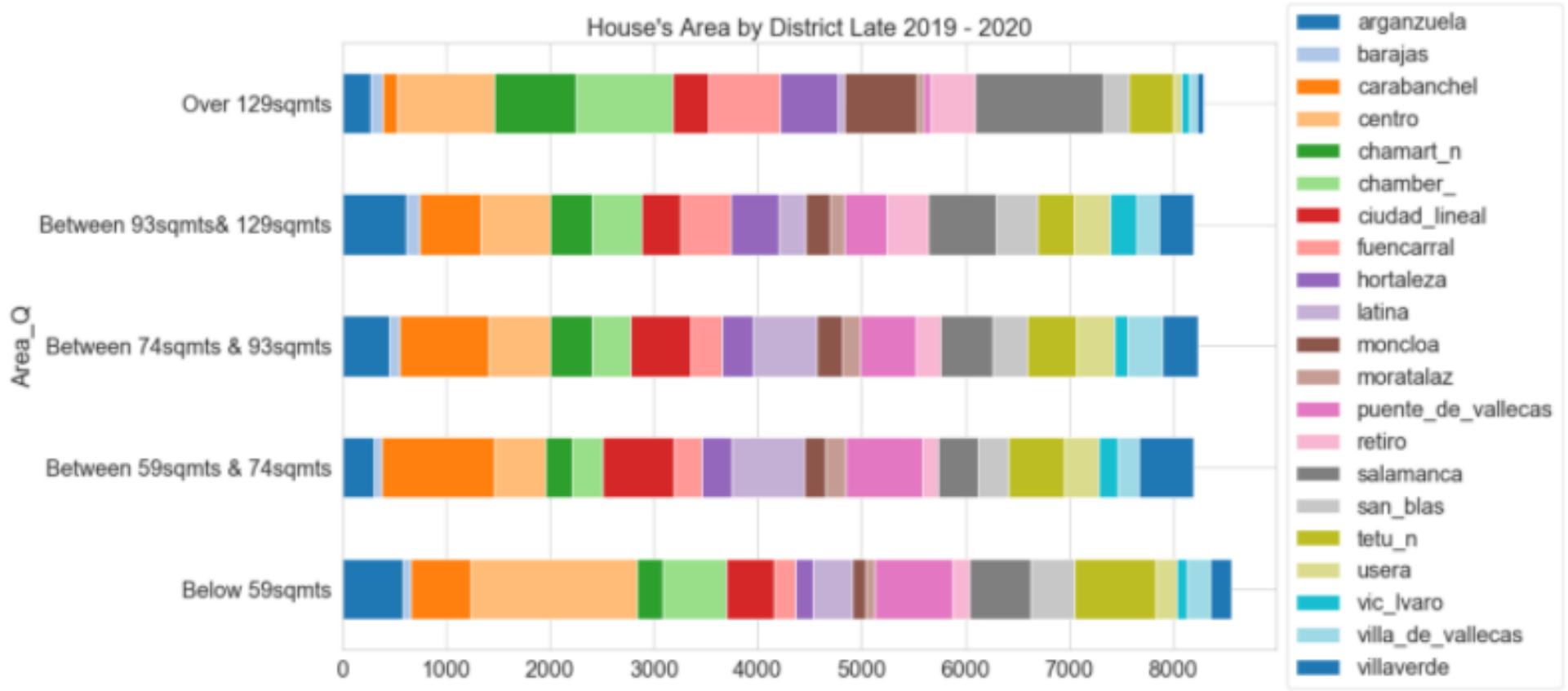

Price_Sq_K_sqm
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2.0
2.0
2.0
2.0
2.0
2.0
2.0
3.0
3.0
3.0
3.0
4.0
4.0
4.0
4.0
4.0
5.0
5.0
5.0
5.0
6.0
villaverde
carabanchel
vic_lvaro
usera
puente_de_vallecas
latina
villa_de_vallecas
barajas
ciudad_lineal
san_blas
moratalaz
tetu_n
arganzuela
fuencarral
hortaleza
moncloa
retiro
chamber_
chamart_n
centro
salamanca
District


House listing in Madrid Late 2019 - 2020
Scrap_date
20191004
20191129
20200105
20200129
20200320
20200628
20200822
20201021
20201202
0
2000
4000
6000
8000
10000
Counter


Houses Listing in Madrid by Districts Late 2019 - 2020
centro
salamanca
carabanchel
chamber_
tetu_n
puente_de_vallecas
ciudad_lineal
arganzuela
chamart_n
latina
fuencarral
hortaleza
Districts
0
1000
2000
3000
4000
Counter


House Pricing by District Late 2019 - 2020
Price_Q
Over 790K
Less than 135K
Between 470K & 790K
Between 329K & 470K
Between 240K & 329K
Between 178K & 240K
Between 135K & 178K
0
1000
2000
3000
4000
5000
6000
arganzuela
carabanchel
centro
chamart_n
chamber_
ciudad_lineal
fuencarral
hortaleza
latina
moncloa
moratalaz
puente_de_vallecas
retiro
salamanca
san_blas
tetu_n
usera
vic_lvaro
villa_de_vallecas
villaverde

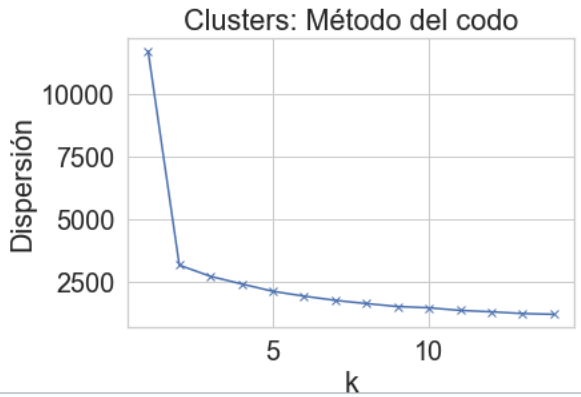

House Price-Sqmts by District Late 2019 - 2020
Price_Q_sqmts
Over 5K
Between 3K & 5K
Between 2K & 3K
Below 2K
0
2000
4000
6000
8000
10000
12000
14000
16000
arganzuela
carabanchel
centro
chamart_n
chamber_
ciudad_lineal
fuencarral
hortaleza
latina
moncloa
moratalaz
puente_de_vallecas
retiro
salamanca
san_blas
tetu_n
usera
vic_lvaro
villa_de_vallecas
villaverde


House's Area by District Late 2019 - 2020
Area_Q
Over 129sqmts
Between 93sqmts& 129sqmts
Between 74sqmts & 93sqmts
Between 59sqmts & 74sqmts
Below 59sqmts
0
1000
2000
3000
4000
5000
6000
7000
8000
arganzuela
barajas
carabanchel
centro
chamart_n
chamber_
ciudad_lineal
fuencarral
hortaleza
latina
moncloa
moratalaz
puente_de_vallecas
retiro
salamanca
san_blas
tetu_n
usera
vic_lvaro
villa_de_vallecas
villaverde

In [94]:

```
#Identificando el numero de clusters
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

#Metodo codo
max_k=15
inertia = []
for k in range(1, max_k):
    kmeans = KMeans(n_clusters = k).fit(scaler_X)
    inertia.append(kmeans.inertia_)

plt.plot(range(1, max_k), inertia, 'bx-')
plt.xlabel('k')
plt.ylabel(u'Dispersión')
plt.title("Clusters: Método del codo")
```

Out[94]: Text(0.5, 1.0, 'Clusters: Método del codo')



In [99]:

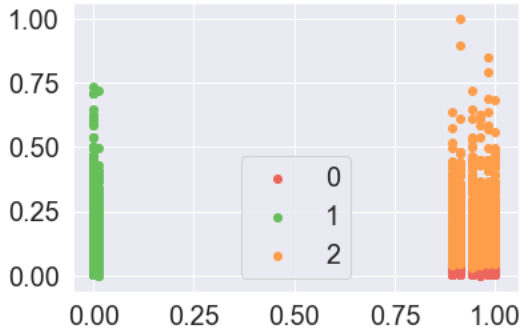
```
kmeans = KMeans(n_clusters=3, random_state=0)
label = kmeans.fit_predict(scaler_X)
print(len(kmeans.labels_))
print(label)
```

```
41467
[1 1 1 ... 0 2 2]
```

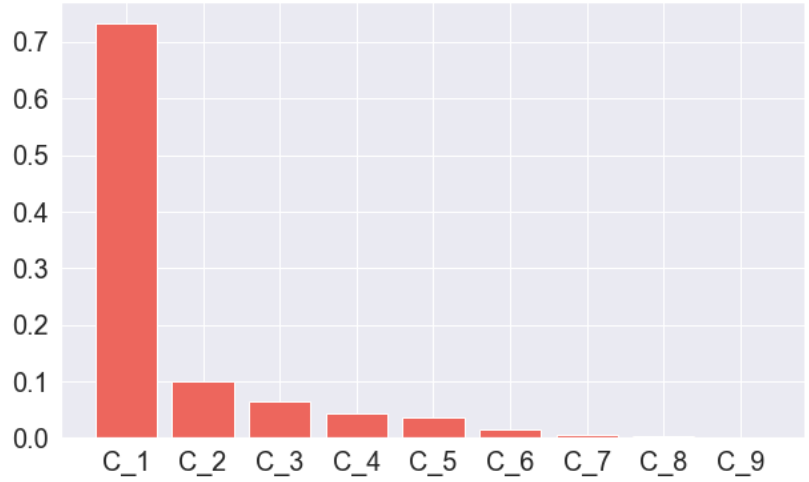
In [102]:

```
#Getting unique labels

u_labels = np.unique(label)
print(u_labels)
#plotting the results:

for i in u_labels:
    plt.scatter(scaler_X[label == i , 0] , scaler_X[label == i , 1] , label = i)
plt.legend()
plt.show()
```

```
[0 1 2]
```



In [104]:
```
fig, ax = plt.subplots(figsize=[10, 6])
ax.bar(
    [f"C_{i+1}" for i in range(dfpca.explained_variance_ratio_.size)],
    dfpca.explained_variance_ratio_
)
plt.show()
```

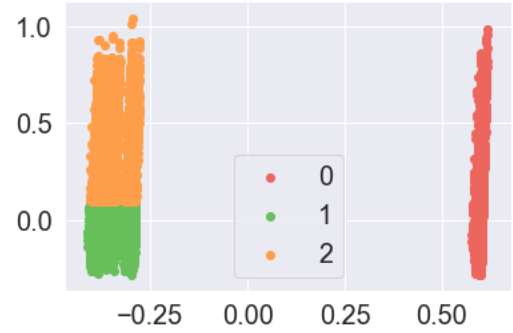


In [105]:
```
for feature, coef in zip(VAR_NUM_2CHECK, dfpca.components_[0]): #va por filas y vas viendo
  print(f"{feature}: {coef:.3f}")
```

```
Scrap_date: -1.000
Area: 0.004
Price_M: 0.015
Year: 0.013
Rooms_num: 0.003
Floor_num: 0.000
Garage_price: -0.000
```

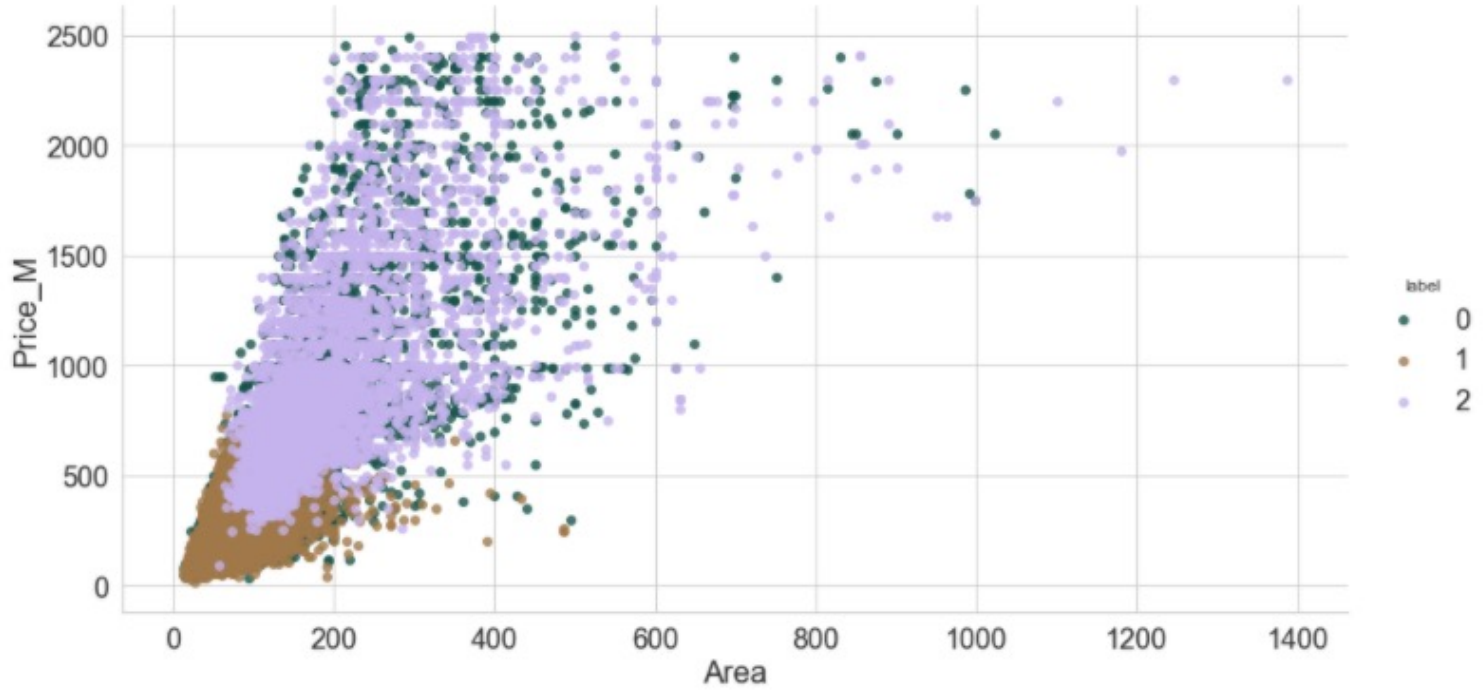
In [110]:
```
#Getting unique labels

u_labels = np.unique(label)
print(u_labels)
#plotting the results:

for i in u_labels:
    plt.scatter(df_pca2_comp[label == i , 0] , df_pca2_comp[label == i , 1] , label = i)
plt.legend()
plt.show()
```

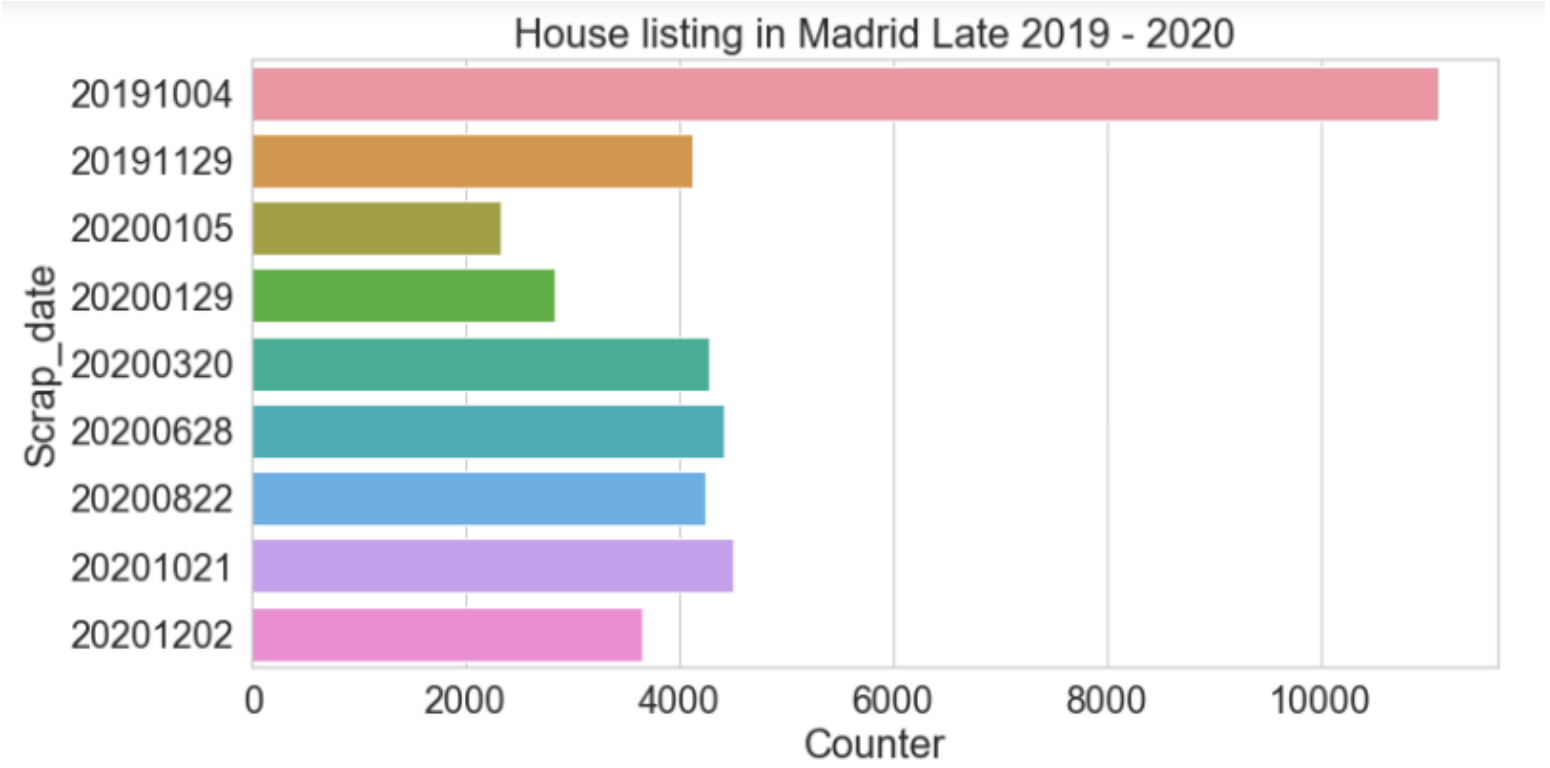
```
[0 1 2]
```



In [115]:

```
sns.set_style('whitegrid')
g=sns.lmplot('Area','Price_M',data=df3_num, hue='label',
             palette='cubehelix',size=12,aspect=1,fit_reg=False)
g.fig.set_figwidth(15.27)
g.fig.set_figheight(6.7)
```




House listing in Madrid Late 2019 - 2020
Scrap_date
20191004
20191129
20200105
20200129
20200320
20200628
20200822
20201021
20201202
0
2000
4000
6000
8000
10000
Counter

PROCESO DEL PROYECTO
MACHINE LEARNING