

# Mid-bootcamp project

Regression



Ignacio Zubia Muñoz  
Pablo Ferrer Coto

- 01 Intro
- 02 Data Cleaning
- 03 SQL
- 04 Power BI
- 05 Conclusiones

# 01 Intro

02 Data Cleaning

03 SQL

04 Power BI

05 Conclusiones

# Intro

---

El objetivo del proyecto es predecir el precio de las viviendas en el estado de Seattle mediante un modelo de regresión lineal.

El conjunto de datos utilizado, contiene información sobre diferentes variables de un elevado número de viviendas del estado de Seattle, incluido el precio de la propiedad, el número de habitaciones, el número de baños, el año de construcción de la propiedad, etc.

Utilizaremos este conjunto de datos para entrenar nuestro modelo de regresión lineal y, a continuación, utilizaremos el modelo para predecir los precios de las propiedades en el conjunto de prueba.

Por otro lado, utilizaremos las funcionalidades de PowerBI para relacionar variables y sacar conclusiones valiosas para el negocio de la compañía



**Dataset**



**Consulta del dataset**



**Data Cleaning y estimación**



**Análisis y conclusiones**

01 Intro

**02 Data Cleaning**

03 SQL

04 Power BI

05 Conclusiones

# Data Cleaning

---

A partir de la limpieza y los modelos estimados, finalmente hemos obtenido un  $R^2 = 0,87$  a partir del modelo Random Forest Regressor

## 01 Importar datos y librerías

-Librerías importadas:

- Numpy
- Pandas
- Seaborn
- Matplotlib
- Datetime

-Dataset importado: Regression.xlsx



## 02 Limpieza de datos

-Eliminamos columnas  
innecesarias: id

-Comprobamos que no hay valores  
NULL

-Comprobamos y eliminamos  
outliers de: bedrooms &  
bathrooms

-Cambiamos el formato de date




## 03 Estimación y validación del modelo

-Regresión Lineal:

- $R^2 = 0,69$
- MAE = 111.111

-Random Forest Regressor:

- $R^2 = 0,87$
  - MAE = 63.432
- 

01 Intro

02 Data Cleaning

**03 SQL**

04 Power BI

05 Conclusiones

# SQL

---

Hemos seguido los pasos necesarios para dar respuesta a la parte de las preguntas de SQL propuestas, trabajando en Jupyter notebook y estableciendo la conexión con SQL.

## Problemas con los que nos hemos encontrado:

```
%%sql
select grade, avg(`condition`) condition_avg
from house_price_data
group by grade
order by grade
```

```
%%sql
select *
from house_price_data
where bedrooms in (3,4)
and bathrooms >3
and floors =1
and waterfront=0
and `condition`>=3
and grade >=5
and price <300000
```

```
%%sql
select * from ( select *, row_number() over
(order by price desc) as price_rank
from house_price_data ) as sub1
where price_rank = 11
```



01 Intro

02 Data Cleaning

03 SQL

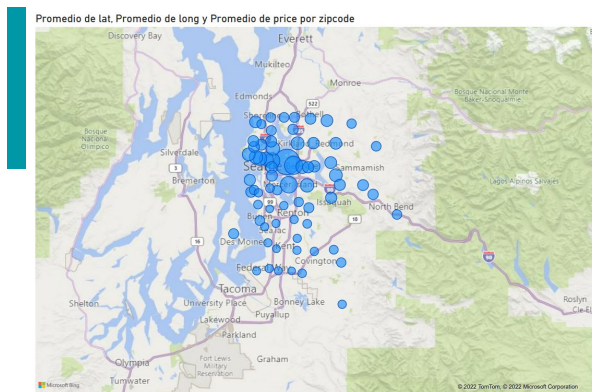
**04 Power BI**

05 Conclusiones

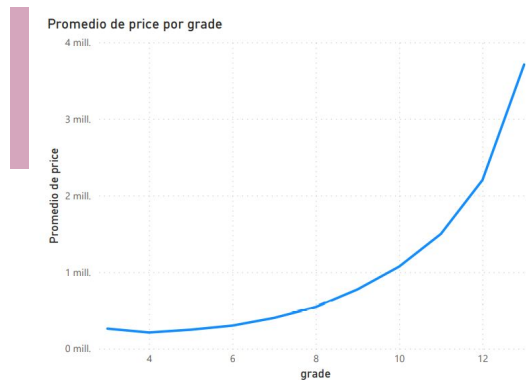
# PowerBI

En esta parte hemos seguido los pasos que se nos habían propuesto para la obtención de los distintos resultados a través de la contestación de las preguntas utilizando el data set contenido en el archivo de regression.xls

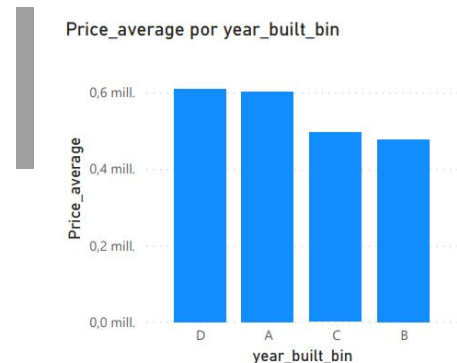
## Resultados a destacar:



La obtención del mapa nos permite situar tanto las viviendas contenidas en el data set como conocer la zona en la que se encuentran



Este gráfico que relaciona el precio promedio de las viviendas con el grado en el que se encuentran estas, explica y relaciona de manera muy visual la relación existente entre estas variables



A destacar de este gráfico es que las viviendas que componen el grupo A (1900-1930) tienen un precio superior que el de los grupos B y C que son de viviendas con un año de construcción más reciente

01 Intro

02 Data Cleaning

03 SQL

04 Power BI

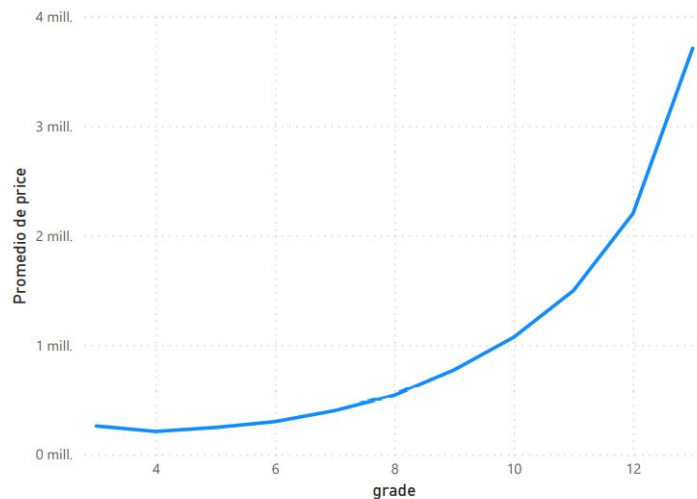
**05 Conclusiones**

# Conclusiones

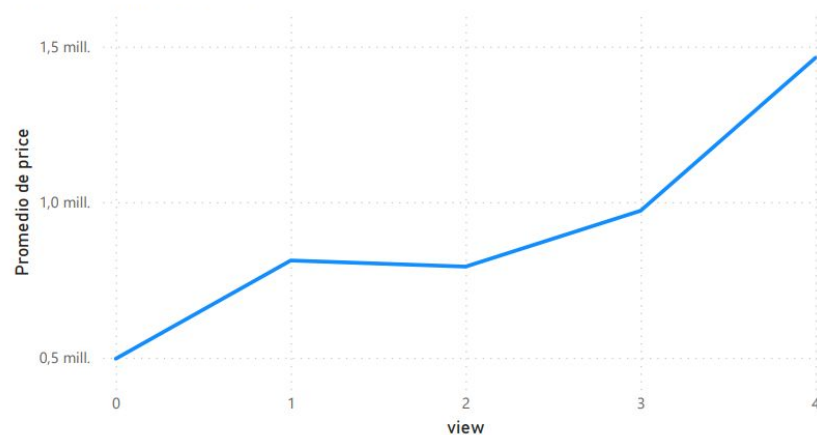
---

Variables con una mayor relación lineal

Grade vs Promedio Price



View vs Promedio Price

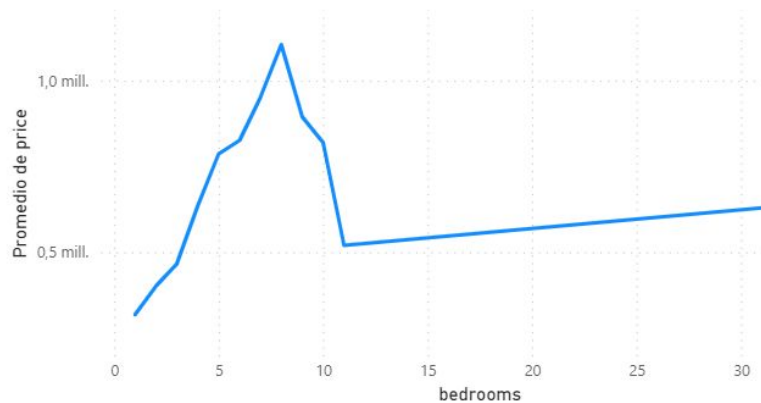


# Conclusiones

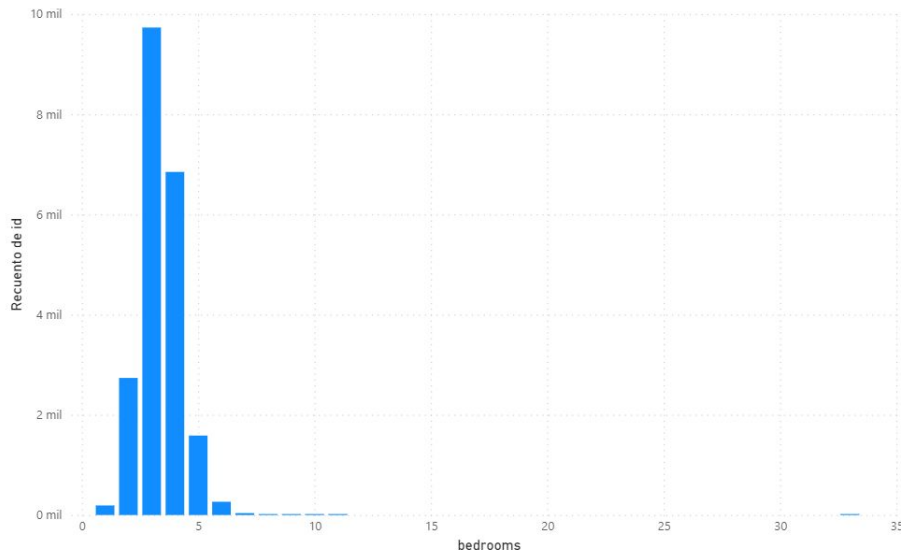
---

Variable Bedroom

**Bedrooms vs Promedio Price**



**Bedrooms vs N° viviendas vendidas**

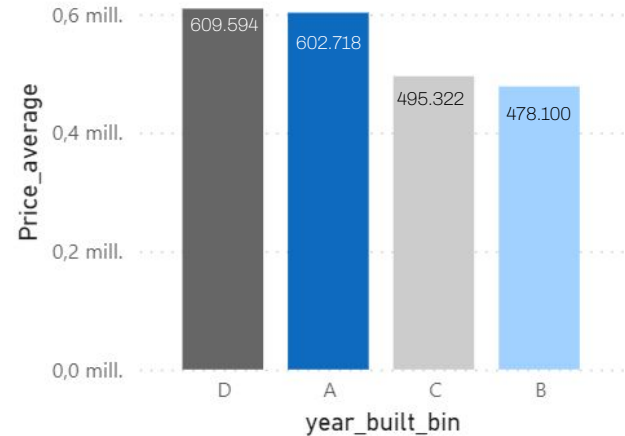


# Conclusiones

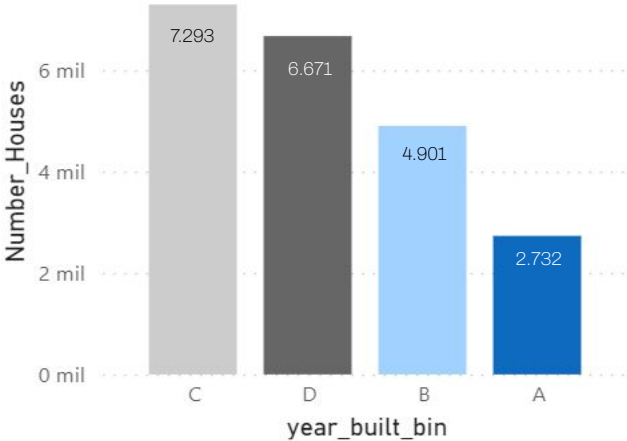
Grupos de Viviendas

<u>Grupos</u>	<u>Años</u>
A	1900 - 30
B	1930 - 60
C	1960 - 90
D	1990 - act

Price Average by year built



Number of Houses by year built



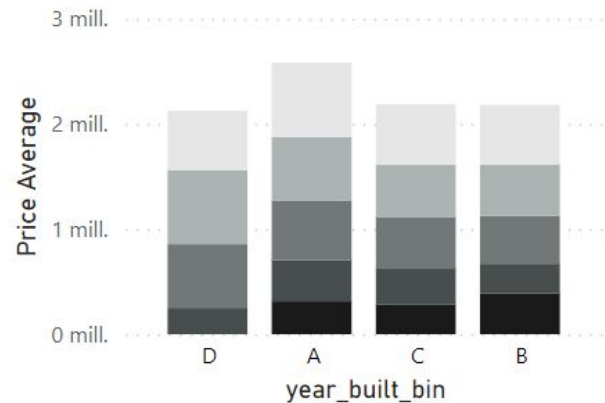
# Conclusiones

Variable condition

**Price Average** by year built with condition filter

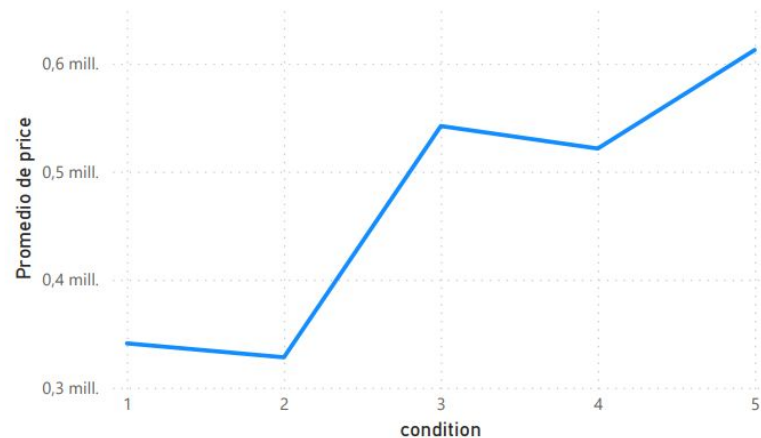
Price Average por year\_built\_bin y condition

condition ● 1 ● 2 ● 3 ● 4 ● 5



**Price Average** vs condition

Promedio de price por condition



Thank you

