California Housing Dataset

This dataset is available both in sklearn package (fetching approach) and also in Kaggle repository, and I personally chose the one from Kaggle.

Source: https://www.kaggle.com/datasets/camnugent/california-housing-prices

The California housing dataset has 20,640 samples and 10 features. 9 of the columns including 'target' are numerical and their dtype is float, and 'ocean Proximity' is the only categorical column. The features and their descriptions are listed below:

1. **longitude**: A measure of how far west a house is; a higher value is farther west
2. **latitude**: A measure of how far north a house is; a higher value is farther north
3. **housingMedianAge**: Median age of a house within a block; a lower number is a newer building
4. **totalRooms**: Total number of rooms within a block
5. **totalBedrooms**: Total number of bedrooms within a block
6. **population**: Total number of people residing within a block
7. **households**: Total number of households, a group of people residing within a home unit, for a block
8. **medianIncome**: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. **medianHouseValue**: Median house value for households within a block (measured in US Dollars)
10. **oceanProximity**: Location of the house w.r.t ocean/sea

Statistical summery

Statistical parameters of the numerical features including mean, standard deviation, minimum, the first quartile (25^{th} percentile), the second quartile or median, the third quartile (75^{th} percentile) and maximum are represented in the below table.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640	20640	20640	20640	20433	20640	20640	20640	20640
mean	-119.57	35.63	28.64	2635.76	537.87	1425.48	499.54	3.87	206855.82
std	003532	2.14	12.59	2181.62	421.39	1132.46	382.33	1.90	115395.62
min	-124.35	32.54	1	2	1	3	1	0.50	14999
25%	-121.8	33.93	18	1447.75	296	787	280	2.56	119600
50%	-118.49	34.26	29	2127	435	1166	409	3.53	179700
75%	-118.01	37.71	37	3148	647	1725	605	4.74	264725
max	-114.31	41.95	52	39320	6445	35682	6082	15	500001

Statistical parameters of the categorical column includes unique values, the value with the most occurrence or Mode, and frequency.

Ocean proximity	
Count	20640
Unique value	5
Top (Mode)	<1H OCEAN
Frequency	9136

Task

Target in this dataset is 'median house value', and it is numeric and continues, because of that regression is the best option to predict the median of the value house and train our model with it.

Challenge

1. make a copy of original data frame in case of wrong modifications and avoid starting again from beginning
2. categorical handling and dealing with classes with too low data points
3. choosing the best way for handling outliers

Concrete Strength Dataset

This dataset is downloaded from Kaggle repository.

Source: https://www.kaggle.com/datasets/mchilamwar/predict-concrete-strength/data

The Concrete Strength dataset has 1,030 samples and 9 columns which one them is target, so this dataset has 8 features. All of the columns are numerical and float except 'AgeInDays' column which is integer. The features and their descriptions are listed below:

1. **CementComponent:** Amount of cement is mixed
2. **BlastFurnaceSlag:** Amount of Blast Furnace Slag is mixed
3. **FlyAshComponent:** Amount of FlyAsh is mixed
4. **WaterComponent:** Amount of water is mixed
5. **SuperplasticizerComponent:** Amount of Super plasticizer is mixed
6. **CoarseAggregateComponent:** Amount of Coarse Aggregate is mixed
7. **FineAggregateComponent:** Amount of Coarse Aggregate is mixed
8. **AgeInDays:** How many days it was left dry
9. **Strength:** What was the final strength of concrete

Statistical summery

Statistical parameters of the numerical features including mean, standard deviation, minimum, the first quartile (25^{th} percentile), the second quartile or median, the third quartile (75^{th} percentile) and maximum are represented in the below table.

	Cement Component	BlastFurnace Slag	FlyAsh Component	Water Component	Superplasticizer Component	CoarseAggregate Component	FineAggregate Component	AgeIn Days	Strength
count	1030.00	1030.00	1030.00	1030.00	1030.00	1030.00	1030.00	1030.00	1030.00
mean	281.17	73.90	54.19	181.57	6.20	972.92	773.58	45.66	35.82
std	104.51	86.28	64.00	21.35	5.97	77.75	80.18	63.17	16.71
min	102.00	0.00	0.00	121.80	0.00	801.00	594.00	1.00	2.33
25%	192.38	0.00	0.00	164.90	0.00	932.00	730.95	7.00	23.71
50%	272.90	22.00	0.00	185.00	6.40	968.00	779.50	28.00	34.45
75%	350.00	142.95	118.30	192.00	10.20	1029.40	824.00	56.00	46.14
max	540.00	359.40	200.10	247.00	32.20	1145.00	992.60	365.00	82.60

Task

I looked for a dataset in which its target was numerical to do regression task on it. After searching in different websites, finally I found this dataset more interesting since my thesis in Bachelor's Degree was about cement. For this reason I chose this one.

Challenge

1. In 3 columns the number of zero values was so high which shows different experiment of adding different amount of each component to test how it changes the strength of cement
2. There was high correlation between Water-Component and Super-plasticizer-Component. Super plasticizer component are high-range water reducers that enhance the workability of concrete without increasing the water content. Dropping one of them should be investigated on the end result to check which one works better.
3. The outlier was only high in AgeinDays column, I used both 'log1p' and 'sqrt' method on the column and the end result of them was equal.

Red Wine Quality

This dataset is downloaded from UCI repository.

Source: https://archive.ics.uci.edu/dataset/186/wine+quality

Red Wine quality dataset has 1,599 samples and 12 columns which one column is the target 'quality'. All the columns are numerical. However, the target column is a discrete numerical value from 1 to 10 and we consider it categorical. The features and their descriptions are listed below:

1. **Fixid acidity**: The non-volatile acids in wine, primarily tartaric acid.
2. **Volatile acidity:** The amount of acetic acid in wine, which at high levels can lead to an unpleasant vinegar taste.
3. **Citric acid:** A minor acid in wine that can add a fresh, slightly citrus flavor.
4. **Residual sugar:** The sugar remaining after fermentation stops, contributing to sweetness.
5. **Chlorides:** The amount of salt in the wine, which can affect taste and preservation.
6. **Free sulfur dioxide:** The amount of SO_2 that is not bound and acts as an antimicrobial and antioxidant.
7. **Total sulfur dioxide:** The total amount of SO_2 present, both free and bound, influencing preservation and potential for spoilage.
8. **Density:** The wine's mass per unit volume, correlating with alcohol and sugar content.
9. **pH:** The level of acidity in the wine, affecting taste and stability.
10. **Sulphates:** Contributes to wine's bitterness and acts as a preservative.
11. Alcohol quality: The ethanol content in wine, impacting body, taste, and aroma.

Statistical summery

Statistical parameters of the numerical features including mean, standard deviation, minimum, the first quartile (25^{th} percentile), the second quartile or median, the third quartile (75^{th} percentile) and maximum are represented in the below table.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599	1599	1599	1599	1599	1599	1599	1.60E+03	1599	1599	1599	1599
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	9.97E-01	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.9	1.89E-03	0.15	0.17	1.07	0.81
min	4.6	0.12	0	0.9	0.01	1	6	9.90E-01	2.74	0.33	8.4	3
25%	7.1	0.39	0.09	1.9	0.07	7	22	9.96E-01	3.21	0.55	9.5	5
50%	7.9	0.52	0.26	2.2	0.08	14	38	9.97E-01	3.31	0.62	10.2	6
75%	9.2	0.64	0.42	2.6	0.09	21	62	9.98E-01	3.4	0.73	11.1	6
max	15.9	1.58	1	15.5	0.61	72	289	1.00E+00	4.01	2	14.9	8

Task

Target in this dataset is 'quality, although it's a numerical variable but it's discrete and limited in the range of 1 to 10. Because of that classification is the best option to predict the quality of the wine and train our model with it.

I wanted to work on a dataset which is related to chemistry since my field of study is chemical engineering, and I found this topic related and also interesting.

Challenge

1. Reading input file with pandas since the file was separated with semicolon
2. High skewness of some columns and I had to test different methods to lower the outliers percentage in that columns which became acceptable to use them in our model
3. Change the label column into binary classification by choosing a limit of 7 which shows a good quality of wine.