Project​​​ Title:​​​ Housing​​​ Pricing​​​ Analysis  
Team​​​ Members:​​​ Edward​​​ Wynman

Dataset​​​ Link:​​​ [Housing​​​ Prices​​​ Dataset​​​ (kaggle.com)](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data)

# Description

* 1. Basic​​​ Information

This​​​ ​project​​​ aims​​​ to​​​ dive​​​ ​into​​​ the​​​ dynamics​​​ of​​​ ​housing​​​ prices​​​ within​​​ the​​​ ​contemporary​​​ real​​​ estate​​​ market​​​ ​of​​​ the​​​ Northeast.​​​ ​The​​​ primary​​​ focus​​​ ​involves​​​ understanding​​​ the​​​ ​influence​​​ of​​​ ​evolving​​​ socio-economic​​​ factors​​​ ​and​​​ recent​​​ ​events​​​ on​​​ ​property​​​ values​​​ ​and​​​ ​​seeing​​​ ​how​​​ ​they​​​ affect​​​ ​housing​​​ prices.​​​

* 1. Project​​​ Objectives​​​

​The​​​ project​​​ revolves​​​ around​​​ ​comprehensive​​​ data​​​ ​exploration,​​​ ​encompassing​​​ data​​​ ​cleaning,​​​ ​detailed​​​ analysis,​​​ ​and​​​ visual​​​ ​representation,​​​ ​seeking​​​ answers​​​ ​to​​​ key​​​ ​inquiries:

* ​Variation​​​ of​​​ house​​​ ​prices​​​ concerning​​​ the​​​ number​​​ of​​​ bedrooms.
* Variation​​​ of​​​ house​​​ prices​​​ concerning​​​ the​​​ number​​​ of​​​ bathrooms.
* Variation​​​ of​​​ house​​​ prices​​​ concerning​​​ the​​​ presence​​​ of​​​ a​​​ basement.
* Variation​​​ of​​​ house​​​ prices​​​ concerning​​​ the​​​ presence​​​ of​​​ air​​​ conditioning.
* Variation​​​ of​​​ house​​​ prices​​​ concerning​​​ the​​​ presence​​​ of​​​ a​​​ guest​​​ room.
  1. Description​​​ of​​​ the​​​ Dataset

​The​​​ housing​​​ price​​​ analysis​​​ project​​​ centered​​​ ​around​​​ a​​​ meticulously​​​ ​compiled​​​ dataset​​​ ​sourced​​​ ​from​​​ ​Kaggle,​​​ ​contributed​​​ by​​​ Yasser​​​ H.​​​ ​This​​​ dataset​​​ presents​​​ a​​​ ​comprehensive​​​ collection​​​ of​​​ ​attributes​​​ vital​​​ for​​​ ​comprehending​​​ housing​​​ price​​​ trends​​​ ​within​​​ the​​​ real​​​ estate​​​ market.​​​ The​​​ ​dataset​​​ encompasses​​​ ​a​​​ diverse​​​ array​​​ of​​​ ​esse​ntial​​​ housing​​​ features​,​​​ offering​​​ ​a​​​ ​holistic​​​​ view​​​ of​​​ the​​​ housing​​​ market​.​​​ ​Key​​​ ​attributes​​​ within​​​ ​the​​​ dataset​​​ ​include​​​ 'price',​​​ 'area',​​​ 'bedrooms',​​​ 'bathrooms',​​​ 'stories',​​​ and​​​ various​​​ binary​​​ indicators​​​ such​​​ as​​​ 'mainroad',​​​ 'guestroom',​​​ 'basement',​​​ 'hotwaterheating',​​​ 'airconditioning',​​​ and​​​ 'parking'.​​​ These​​​ attributes​​​ collectively​​​ provide​​​ an​​​ extensive​​​ ​overview​​​ of​​​ ​housing​​​ characteristics,​​​ ​enabling​​​ nuanced​​​ analysis​​​ ​and​​​ exploration.​​​

​This​​​ dataset​​​ serves​​​ as​​​ ​a​​​ foundational​​​ resource​​​ for​​​ investigating​​​ the​​​ intricate​​​ interplay​​​ between​​​ diverse​​​ housing​​​ attributes​​​ and​​​ ​their​​​ impact​​​ on​​​ housing​​​ prices.​​​ ​Leveraging​​​ this​​​ rich​​​ dataset,​​​ the​​​ project​​​ underwent​​​ ​meticulous​​​ data​​​ preparation,​​​ exploration,​​​ and​​​ analysis,​​​ utilizing​​​ advanced​​​ data​​​ science​​​ techniques​​​ to​​​ derive​​​ meaningful​​​ ​insights​​​ into​​​ the​​​ real​​​ estate​​​ market​​​ dynamics.

# Exploration​​​ of​​​ Data​​​ Analysis

2.1​​​ Data​​​ Preparation

The​​​ i​nitial​​​ phase​​​ of​​​ this​​​ data​​​ ​science​​​ project​​​ involved​​​ ​the​​​ loading​​​ and​​​ preparation​​​ of​​​ the​​​ dataset,​​​ integrating​​​ crucial​​​ ​libraries​​​ such​​​ as​​​ Pandas,​​​ a​​​ ​powerful​​​ data​​​ manipulation​​​ tool.​​​ Leveraging​​​ Pandas,​​​ the​​​ ​dataset​​​ was​​​ loaded​​​ from​​​ the​​​ specified​​​ path​​​ using​​​ the​​​ pd.read\_csv()​​​ function.​​​ This​​​ facilitated​​​​ a​​​ seamless​​​ transition​​​ of​​​ the​​​ dataset​​​ into​​​ a​​​ Pandas​​​ DataFrame,​​​ enabling​​​ further​​​ exploration​​​ and​​​ ​manipulation.

The​​​ data​​​ cleaning​​​ ​process​​​ ensued,​​​ focusing​​​ on​​​ enhancing​​​ the​​​ dataset's​​​ quality​​​ and​​​ completeness.​​​ Essential​​​ ​columns​​​ vital​​​ for​​​ the​​​ housing​​​ price​​​ analysis,​​​ namely​​​ 'price',​​​ 'area',​​​ 'bedrooms',​​​ 'bathrooms',​​​ 'stories',​​​ 'mainroad',​​​ 'guestroom',​​​ 'basement',​​​ 'hotwaterheating',​​​ 'airconditioning',​​​ and​​​ 'parking',​​​ were​​​ ​selectively​​​ retained.​​​ These​​​ columns​​​ were​​​ identified​​​ as​​​ instrumental​​​ for​​​ addressing​​​ the​​​​ research​​​ questions​​​ set​​​ forth​​​ in​​​ the​​​ project's​​​ objectives.

Some​​​ of​​​ the​​​ data​​​ was​​​ ​represented​​​ as​​​ 'yes'​​​ and​​​ 'no'​​​ within​​​ specific​​​ columns,​​​ signifying​​​ binary​​​ attributes​​​ such​​​ as​​​ 'mainroad',​​​ 'guestroom',​​​ 'basement',​​​ 'hotwaterheating',​​​ and​​​ 'airconditioning',​​​ were​​​ meticulously​​​ ​transformed​​​ into​​​ a​​​ numerical​​​ format.​​​ This​​​ conversion,​​​ achieved​​​ using​​​ Pandas'​​​ efficient​​​ applymap()​​​ function,​​​ ​encoded​​​ 'yes'​​​ as​​​ 1​​​ and​​​ 'no'​​​ as​​​ 0.​​​ This​​​ transformation​​​ facilitated​​​ the​​​ ease​​​ of​​​ computation​​​ and​​​ analysis​​​ ​while​​​ ensuring​​​ consistency​​​ within​​​ the​​​ dataset.​​​ To​​​ make​​​ the​​​ analysis​​​ process​​​ easier​​​ and​​​ to​​​ remove​​​ ​redundant​​​ or​​​ non-contributory​​​ information,​​​ the​​​ 'furnishingstatus'​​​ column​​​ was​​​​ dropped.​​​ This​​​ step​​​ was​​​ undertaken​​​ after​​​ careful​​​ consideration​​​ of​​​ the​​​ column's​​​ relevance​​​ to​​​ the​​​ ​analysis​​​ objectives​​​ and​​​ the​​​ project's​​​ scope.

In​​​ addition​​​ to​​​ column​​​ selection​​​ and​​​ transformation,​​​ special​​​ attention​​​ was​​​ given​​​ to​​​ handling​​​ missing​​​ values​​​ within​​​ the​​​ 'area'​​​ column.​​​ Imputation​​​ techniques​​​ were​​​ employed​​​ to​​​ replace​​​ these​​​ missing​​​ values​​​ with​​​ the​​​ median​​​ value​​​ of​​​ the​​​ 'area'​​​ column,​​​ ensuring​​​ the​​​ dataset's​​​ completeness​​​ and​​​ accuracy​​​ in​​​ subsequent​​​ analyses.

2.2​​​ Data​​​ Analysis​​​ using​​​ descriptive​​​ statistics.

The​​​ analysis​​​ employed​​​ descriptive​​​ statistics​​​ to​​​ unravel​​​ pivotal​​​ insights​​​ hidden​​​ within​​​ the​​​ dataset's​​​ numeric​​​ columns.​​​ This​​​ phase​​​ of​​​ the​​​ project​​​ heavily​​​ relied​​​ on​​​ essential​​​ libraries​​​ like​​​ Pandas​​​ for​​​ data​​​ manipulation,​​​ Streamlit​​​ for​​​ visualization,​​​ and​​​ Matplotlib​​​ and​​​ Seaborn​​​ for​​​ graphical​​​ representation.​​​ Utilizing​​​ these​​​ libraries,​​​ the​​​ exploration​​​ centered​​​ on​​​ elucidating​​​ key​​​ statistical​​​ metrics,​​​ showcasing​​​ the​​​ mean​​​ prices​​​ associated​​​ with​​​ distinctive​​​ housing​​​ attributes.​​​ Specifically,​​​ the​​​ impact​​​ of​​​ various​​​ factors​​​ like​​​ the​​​ presence​​​ of​​​ a​​​ basement,​​​ air​​​ conditioning,​​​ and​​​ guest​​​ room​​​ availability​​​ on​​​ house​​​ prices​​​ was​​​ scrutinized.​​​

This​​​ involved​​​ calculating​​​ mean​​​ prices​​​ across​​​ different​​​ categories​​​ of​​​ these​​​ attributes,​​​ providing​​​ a​​​ comprehensive​​​ understanding​​​ of​​​ their​​​ influence​​​ on​​​ the​​​ overall​​​ housing​​​ prices.​​​ The​​​ implementation​​​ of​​​ Pandas​​​ facilitated​​​ efficient​​​ data​​​ handling,​​​ enabling​​​ the​​​ extraction​​​ of​​​ relevant​​​ subsets​​​ based​​​ on​​​ specific​​​ attribute​​​ criteria.​​​ Additionally,​​​ the​​​ use​​​ of​​​ Streamlit​​​ enhanced​​​ the​​​ visualization​​​ process,​​​ empowering​​​ the​​​ creation​​​ of​​​ intuitive​​​ and​​​ interactive​​​ graphs.​​​ Matplotlib​​​ and​​​ Seaborn​​​ played​​​ pivotal​​​ roles​​​ in​​​ crafting​​​ informative​​​ visual​​​ representations,​​​ ranging​​​ from​​​ line​​​ graphs​​​ to​​​ elucidate​​​ mean​​​ prices​​​ concerning​​​ specific​​​ housing​​​ features​​​ to​​​ bar​​​ plots​​​ showcasing​​​ the​​​ distribution​​​ of​​​ categorical​​​ attributes.​​​

2.3​​​ Other​​​ techniques

Along​​​ with​​​ the​​​ descriptive​​​ statistics​​​ analysis,​​​ other​​​ pivotal​​​ techniques​​​ were​​​ employed​​​ to​​​ unearth​​​ deeper​​​ insights​​​ within​​​ the​​​ dataset.​​​ This​​​ phase​​​ encompassed​​​ the​​​ utilization​​​ of​​​ advanced​​​ data​​​ aggregation​​​ techniques,​​​ leveraging​​​ Pandas​​​ functionalities​​​ extensively.​​​

The​​​ implementation​​​ of​​​ data​​​ aggregation​​​ centered​​​ on​​​ calculating​​​ the​​​ mean​​​ prices​​​ based​​​ on​​​ the​​​ number​​​ of​​​ stories​​​ attributed​​​ to​​​ houses.​​​ This​​​ was​​​ achieved​​​ by​​​ grouping​​​ the​​​ dataset​​​ based​​​ on​​​ the​​​ 'stories'​​​ column​​​ and​​​ computing​​​ the​​​ mean​​​ price​​​ within​​​ each​​​ group.​​​ Pandas'​​​ groupby()​​​ functionality​​​ was​​​ instrumental​​​ in​​​ this​​​ process,​​​ allowing​​​ the​​​ aggregation​​​ of​​​ data​​​ according​​​ to​​​ the​​​ number​​​ of​​​ stories,​​​ and​​​ subsequently,​​​ deriving​​​ insightful​​​ mean​​​ price​​​ statistics.​​​ This​​​ technique​​​ contributed​​​ significantly​​​ to​​​ understanding​​​ the​​​ nuanced​​​ impact​​​ of​​​ the​​​ number​​​ of​​​ stories​​​ on​​​ housing​​​ prices,​​​ offering​​​ a​​​ multifaceted​​​ perspective​​​ essential​​​ for​​​ comprehensive​​​ analysis.

# Data​​​ Visualization

The​​​ data​​​ visualization​​​ phase​​​ was​​​ an​​​ integral​​​ part​​​ of​​​ this​​​ project,​​​ aiming​​​ not​​​ just​​​ to​​​ present​​​ numerical​​​ figures​​​ but​​​ to​​​ provide​​​ a​​​ visually​​​ compelling​​​ narrative,​​​ leveraging​​​ Matplotlib,​​​ Seaborn,​​​ and​​​ Streamlit​​​ to​​​ craft​​​ informative​​​ visual​​​ representations.

3.1​​​ Line​​​ Graphs​​​ for​​​ Housing​​​ Prices​​​ by​​​ Number​​​ of​​​ Bedrooms​​​ and​​​ Bathrooms:

The​​​ creation​​​ of​​​ line​​​ graphs​​​ was​​​ instrumental​​​ in​​​ elucidating​​​ the​​​ relationship​​​ between​​​ house​​​ prices​​​ and​​​ the​​​ number​​​ of​​​ bedrooms​​​ and​​​ bathrooms.​​​ These​​​ graphs​​​ were​​​ meticulously​​​ crafted​​​ using​​​ Matplotlib​​​ and​​​ Seaborn​​​ libraries,​​​ ensuring​​​ clarity,​​​ precision,​​​ and​​​ interpretability.​​​ The​​​ line​​​ graphs​​​ were​​​ structured​​​ to​​​ showcase​​​ mean​​​ house​​​ prices​​​ across​​​ different​​​ categories​​​ of​​​ bedrooms​​​ and​​​ bathrooms.​​​ The​​​ x-axis​​​ denoted​​​ the​​​ varying​​​ number​​​ of​​​ bedrooms​​​ or​​​ bathrooms,​​​ while​​​ the​​​ y-axis​​​ depicted​​​ mean​​​ house​​​ prices,​​​ measured​​​ in​​​ millions​​​ of​​​ USD.​​​ Each​​​ plotted​​​ point​​​ on​​​ the​​​ graph​​​ corresponded​​​ to​​​ the​​​ mean​​​ price​​​ associated​​​ with​​​ a​​​ specific​​​ category​​​ of​​​ bedrooms​​​ or​​​ bathrooms.

These​​​ line​​​ graphs​​​ not​​​ only​​​ depicted​​​ the​​​ fluctuation​​​ in​​​ mean​​​ prices​​​ concerning​​​ the​​​ changing​​​ count​​​ of​​​ bedrooms​​​ and​​​ bathrooms​​​ but​​​ also​​​ facilitated​​​ the​​​ identification​​​ of​​​ potential​​​ trends​​​ or​​​ patterns​​​ within​​​ the​​​ dataset.​​​ Clear​​​ labeling​​​ of​​​ axes,​​​ precise​​​ data​​​ points,​​​ and​​​ distinctive​​​ visualization​​​ styles​​​ aided​​​ in​​​ presenting​​​ these​​​ insights​​​ in​​​ an​​​ easily​​​ comprehensible​​​ manner.

3.2​​​ Interactive​​​ and​​​ Informative​​​ Visualizations​​​ Using​​​ Streamlit:

Utilizing​​​ Streamlit​​​ as​​​ the​​​ visualization​​​ platform​​​ added​​​ an​​​ interactive​​​ dimension​​​ to​​​ the​​​ graphical​​​ representations.​​​ Streamlit's​​​ intuitive​​​ functionalities​​​ enabled​​​ the​​​ creation​​​ of​​​ user-friendly​​​ interfaces,​​​ allowing​​​ stakeholders​​​ to​​​ dynamically​​​ interact​​​ with​​​ and​​​ explore​​​ the​​​ data.​​​ The​​​ visualizations​​​ embedded​​​ within​​​ the​​​ Streamlit​​​ app​​​ were​​​ designed​​​ to​​​ be​​​ informative​​​ and​​​ engaging.​​​ They​​​ provided​​​ a​​​ platform​​​ for​​​ users​​​ to​​​ select​​​ specific​​​ parameters,​​​ toggle​​​ between​​​ different​​​ visual​​​ representations,​​​ and​​​ obtain​​​ instantaneous​​​ insights​​​ into​​​ the​​​ fluctuating​​​ trends​​​ of​​​ housing​​​ prices​​​ concerning​​​ various​​​ attributes.​​​ Interactive​​​ features​​​ such​​​ as​​​ tooltips​​​ and​​​ zooming​​​ capabilities​​​ enriched​​​ the​​​ user​​​ experience,​​​ empowering​​​ stakeholders​​​ to​​​ delve​​​ deeper​​​ into​​​ the​​​ dataset​​​ and​​​ extract​​​ meaningful​​​ conclusions​​​ effortlessly.

These​​​ visualizations​​​ not​​​ only​​​ complemented​​​ the​​​ statistical​​​ analysis​​​ but​​​ also​​​ served​​​ as​​​ powerful​​​ communication​​​ tools,​​​ effectively​​​ conveying​​​ complex​​​ trends​​​ and​​​ relationships​​​ within​​​ the​​​ dataset.​​​ They​​​ played​​​ a​​​ crucial​​​ role​​​ in​​​ facilitating​​​ informed​​​ decision-making​​​ and​​​ understanding​​​ the​​​ nuances​​​ of​​​ housing​​​ prices​​​ concerning​​​ different​​​ attributes.

# Data​​​ manipulating

The​​​ data​​​ manipulation​​​ segment​​​ of​​​ this​​​ project​​​ revolved​​​ around​​​ refining​​​ the​​​ dataset,​​​ focusing​​​ on​​​ handling​​​ missing​​​ values​​​ and​​​ performing​​​ strategic​​​ data​​​ aggregation​​​ to​​​ enhance​​​ the​​​ dataset's​​​ completeness​​​ and​​​ richness​​​ for​​​ in-depth​​​ analysis.

4.1​​​ Handling​​​ Missing​​​ Values​​​ in​​​ the​​​ 'Area'​​​ Column:

A​​​ critical​​​ aspect​​​ of​​​ data​​​ manipulation​​​ involved​​​ addressing​​​ missing​​​ values​​​ within​​​ the​​​ 'area'​​​ column,​​​ a​​​ pivotal​​​ attribute​​​ influencing​​​ housing​​​ prices.​​​ The​​​ implementation​​​ of​​​ Pandas'​​​ functionalities​​​ facilitated​​​ a​​​ systematic​​​ approach​​​ to​​​ handle​​​ these​​​ missing​​​ values.​​​ Strategies​​​ employed​​​ for​​​ missing​​​ value​​​ treatment​​​ centered​​​ on​​​ imputing​​​ these​​​ values​​​ with​​​ the​​​ median​​​ of​​​ the​​​ 'area'​​​ column.​​​ This​​​ process​​​ ensured​​​ that​​​ missing​​​ values​​​ were​​​ replaced​​​ with​​​ a​​​ statistically​​​ representative​​​ value,​​​ preserving​​​ the​​​ integrity​​​ of​​​ the​​​ dataset.​​​ This​​​ meticulous​​​ treatment​​​ of​​​ missing​​​ values​​​ contributed​​​ significantly​​​ to​​​ the​​​ completeness​​​ and​​​ accuracy​​​ of​​​ the​​​ dataset,​​​ eliminating​​​ potential​​​ biases​​​ that​​​ could​​​ arise​​​ from​​​ incomplete​​​ data.

4.2​​​ Data​​​ Aggregation​​​ for​​​ Deriving​​​ Insights​​​ Based​​​ on​​​ Stories:

Additionally,​​​ data​​​ aggregation​​​ techniques​​​ were​​​ strategically​​​ utilized​​​ to​​​ derive​​​ insights​​​ based​​​ on​​​ the​​​ number​​​ of​​​ stories​​​ attributed​​​ to​​​ houses.​​​ Leveraging​​​ Pandas'​​​ groupby()​​​ functionality,​​​ the​​​ dataset​​​ was​​​ grouped​​​ according​​​ to​​​ the​​​ 'stories'​​​ column,​​​ enabling​​​ the​​​ calculation​​​ of​​​ mean​​​ prices​​​ within​​​ each​​​ group.​​​ This​​​ process​​​ of​​​ aggregating​​​ data​​​ based​​​ on​​​ the​​​ number​​​ of​​​ stories​​​ allowed​​​ for​​​ the​​​ extraction​​​ of​​​ insightful​​​ statistics.​​​ These​​​ statistics​​​ provided​​​ valuable​​​ information​​​ about​​​ the​​​ correlation​​​ between​​​ the​​​ number​​​ of​​​ stories​​​ and​​​ housing​​​ prices,​​​ offering​​​ a​​​ nuanced​​​ perspective​​​ that​​​ contributed​​​ to​​​ a​​​ deeper​​​ understanding​​​ of​​​ the​​​ dataset.

The​​​ meticulous​​​ handling​​​ of​​​ missing​​​ values​​​ and​​​ the​​​ strategic​​​ implementation​​​ of​​​ data​​​ aggregation​​​ techniques​​​ were​​​ pivotal​​​ steps​​​ in​​​ enhancing​​​ the​​​ dataset's​​​ quality​​​ and​​​ depth.​​​ These​​​ manipulations​​​ ensured​​​ a​​​ more​​​ robust​​​ dataset,​​​ empowering​​​ subsequent​​​ analyses​​​ to​​​ yield​​​ more​​​ accurate​​​ and​​​ comprehensive​​​ insights​​​ into​​​ the​​​ factors​​​ influencing​​​ housing​​​ prices.

# Conclusion

In​​​ conclusion,​​​ the​​​ amalgamation​​​ of​​​ data​​​ exploration,​​​ preparation,​​​ analysis,​​​ visualization,​​​ and​​​ manipulation​​​ presented​​​ a​​​ comprehensive​​​ overview​​​ of​​​ housing​​​ price​​​ dynamics.​​​ The​​​ project's​​​ success​​​ lay​​​ not​​​ only​​​ in​​​ deriving​​​ insights​​​ from​​​ the​​​ dataset​​​ but​​​ also​​​ in​​​ effectively​​​ communicating​​​ these​​​ findings​​​ through​​​ intuitive​​​ visualizations.​​​ The​​​ project's​​​ outcomes​​​ pave​​​ the​​​ way​​​ for​​​ informed​​​ decision-making​​​ within​​​ the​​​ real​​​ estate​​​ domain,​​​ highlighting​​​ the​​​ intrinsic​​​ relationship​​​ between​​​ diverse​​​ housing​​​ attributes​​​ and​​​ their​​​ impact​​​ on​​​ prices.​​​ The​​​ multifaceted​​​ approach​​​ adopted​​​ throughout​​​ this​​​ project​​​ accentuates​​​ the​​​ importance​​​ of​​​ leveraging​​​ advanced​​​ data​​​ science​​​ techniques,​​​ showcasing​​​ their​​​ prowess​​​ in​​​ unraveling​​​ intricate​​​ patterns​​​ within​​​ real​​​ estate​​​ datasets.