Steps for complete DATA SCIENCE / ML project

# Exploratory data analysis

* 1. Dataset information
  2. Generate notes.xlsx: list all features and WRITE their description
  3. Check for missing values & duplicated values
  4. Explore categorical > adjust features (delete or shift to numerical) & write observations
  5. Explore numerical > adjust features (delete or shift to categorical) & write observations

# Data preprocessing

* 1. Handle missing data
  2. Get a new list of categorical / numerical features
  3. Handle outliers
  4. Feature engineering if possible

# Data analysis & storytelling

* 1. Generating lots of questions
  2. Generating answers for questions
  3. Wrapping observations in a wonderful story

# Modelling

* 1. Raw modelling as benchmark and future comparison
  2. Check for clustering
  3. Model numerical features
  4. Feature selection / engineering for numerical features
  5. Add categorical features to my model starting with lower number of unique cats in every feature
  6. Feature selection / engineering for categorical features
  7. Modelling different features against different algorithms and generating cross table
  8. Choosing the optimum model / features and hyperparameters optimization

# Forecasting

* 1. Data preprocessing for modelling
  2. forecasting

**Some questions:**

1. Does property type [MSSubClass] have an effect on [SalePrice]?

* 50% of observation are in category 20 and 60, and they contain most of outliers
* All categories share the same range of SalePrice

1. Relationship between MSZoning and SalePrice

* 93% of observation are on 2 categories RL and RM
* All categories share the same range of SalePrice

1. Relationship between Neighborhood, MSSubClass, MSZoning and SalePrice

* Neighborhoods [NridgHt and StoneBr and NoRidge] have high range of SalePrice and the rest of neighborhoods share the same range.
* MSZoning:
* RL in all neighborhood
* FV only in Somerst
* RM mainly in BrDale,BrkSide,IDOTRR,MeadowVm,OldTown

1. Relationship between LotArea, LotFrontage and SalePrice

* In general, there is no correlation between LotArea, LotFrontage and SalePrice even if we show variations in Neighborhood or MSZoning

1. Relationship between OverallQual and OverallCond

* There is not relation between OverallQual and OverallCond

1. Relationship between OverallQual and OverallCond and YearBuilt and YearRemodAdd

* There is no relation between YearRemodAdd and OverallQual or OverallCond

1. When do they Remod the property

* Alarm: in 53% of data, YearBuilt and YearRemodAdd are the same which is not logic, may be it’s used to describe the last time the property was built/renewed
* Remod period ranges from 20 to 100 years

1. Relationship between YearBuilt / YearRemodAdd and SalePrice

* There is positive correlation between YearRemodAdd and SalePrice
* I think YearRemodAdd will be more representative for SalePrice than YearBuilt I will drop the following columns: YearBuilt / GarageYrBlt since they don't add any new value to our dataset
* Positive correlation with high width between YearBuilt and SalePrice
* Alarm: in 53% of data, YearBuilt and YearRemodAdd are the same which is not logic, may be it’s used to describe the last time the property was built/renewed
* In 60% of data, GarageYearBlt is the same as property yearbuilt which is logic. However, there are 17 observations where garage was built before property

1. Relationship between BsmtFinSF1 and BsntFinSF2 and BsmtUnfSF and TotalBsmtSF, BsmtFullBath, BsmtQual abd BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2

* There is a noticeable positive correlation between TotalBsmtSF and 1stFlrSF and SalePrice

1. Is there any relation between Bsmt and LotArea or SalePrice

* There is positive correlation between BsmtFinSF1 and SalePrice and no correlation to LotArea

1. Do SalePrice affected by month sold over the years

* Most of sales are done in 2nd and 3rd quarter of the year

1. Does Condition1 affect any feature?

* It’s important for our model I think

1. RoofStyle and LotArea or Saleprice

* No relation

1. SaleType and SaleCondition

* Only Salecondition: new are sold partially

1. Does Utilities like: Heating, HeatingQC, CentrailAir, Electrical affect SalePrice, SaleType?

* Heating & HeatingQC: other than GasA, you have a very limited options in saleprice
* CentralAir: if property doesn’t have centralAir, it will not have GasA heating and it’s price will drop dramatically
* Electrical: Saleprice will drop if electrical system is other than SBrkr

1. Can we combine the above utilities feature into a new one?

* I couldn’t