| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Python | Create KaggleKernel account | | | | | | | |
| | | https://colab.research.google.com/notebooks/intro.ipynb#recent=true | | | | | | |
| | Create Notebook - Python | | | | | | | |
| | | | | | | | | |
| | Pre-Add Data | | | | | | | |
| | | https://github.com/gcaseycupp/BeginnerMLOverviewProject/tree/master/Week2 | | | | | | |
| | | Download csv 400Sales | | | | | | |
| | | Rename - 400PreProcessed.csv | | | | | | |
| | Kaggle - Add Data | | | | | | | |
| | Upload | | | | | | | |
| | DatasetTitle - PreProcessed | | | | | | | |
| | If you see red 'Draft Session Fatal' Refresh | | | | | | | |
| | Run notebook, you will likely NOT see files yet.. takes a bit... so let's go work on our data.. this is a Kaggle problem, not a python problem.. i | | | | | | | |
| | | | | | | | | |
| Data | Home Prices | | | | | | | |
| | | http://data.opencolorado.org/dataset/city-and-county-of-denver-real-property-sales-book-2013/resource/2e1e28ba-cb31-4959-bc41-e16e573e8015?inner_span=True | | | | | | |
| | | | | | | | | |
| | 3rd One down is csv | | | | | | | |
| | File-> Import-> Upload -> Grad file from above | | | | | | | |
| | Rename Sheet | | | | | | | |
| | | Rename as Full Data | | | | | | |
| | | Duplicate Sheet (might be slow) | | | | | | |
| | | | | | | | | |
| | Delete Columns | Rename sheet as 11 Columns | | | | | | |
| | | Model | Cluster | Group | CondoComplex | CondoComplexName | | |
| | | ABSTRPRD | | | | | | |
| | | YRBUILT | | | | | | |
| | | TOTUNITS | STORIES | GARTYPE | | | | |
| | | XTRAFFIC | FLOORLVL | | | | | |
| | | COMMUSE | ARCSTYLE | SPRICE | DEDUCT | | | |
| | | TASP | SMONTH | SYEAR | QUALIFIED | STATUS | | |
| | | Duplicate Sheet | | | | | | |
| | | | | | | | | |
| | Reduce Rows | Rename Sheet as Working Data | | | | | | |
| | | https://www.denvergov.org/Maps/map/assessormaps | | | | | | |
| | | Location, Location, Location | | | | | | |
| | | Data -> Create a Filter -> By Condition (or checkmarks) | | | | | | |
| | | By NBHD, starts with 71 | | | | | | |
| | | | | | | | | |
| | | Sort By Numbdrm | | | | | | |
| | | | Select header of column, right click , sort a-z scroll to bottom | | | | | |
| | | | Delete Empty | | | | | |
| | | Sort By living area | | | | | | |
| | | | delete range | | | | | |
| | | Sort by Price | | | | | | |
| | | | Delete 5 million.. | Always Domain Knowledge | | | | |
| | | Delete Parcel Column | | | | | | |
| | | | | | | | | |
| | | Create New Sheet - Don't Duplicate - Copy and Paste (click top left box) | | | | | | |
| | | | | | | | | |
| | Working | Rename Sheet as CreateNewColumn | | | | | | |
| | | Categorical - Data - much harder w/ python - for NBHD, you can't use strings, but it usually has problems w/ numbers.. onecommon is 1-many columns ... getDummies()   holidays | | | | | | |
| | | CONDITION_ADJ | | | | | | |
| | | | Insert Column after Condition | | | | | |

|  |  |  |  | =if(E2="Good",1,0) |
|---|---|---|---|---|
|  |  |  | Duplicate final sheet |  |
|  |  |  |  |  |
|  |  |  | Download CSV |  |
|  |  |  |  |  |
|  | Python | Follow code |  |  |

| 3 | Select the Algorithm |  |  |
|---|---|---|---|
|  |  | Home Prices | Azure |
|  |  |  | https://studio.azureml.net/ |
|  |  |  | Not an Azure ML Studio (classic) user? |

| | | | | | | Sign up here | https://studio.azureml.net/# | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 8 hour (d**oesn't allow dataset upload**) so Free, No Credit Card required | | | | | |
| | | | | | **Middle one that says 'Free' is not Free.. BEWARE sneaky charges** | | | | | |
| | | | Upload Data | | | | | | | |
| | | | | Click Datasets | | | | | | |
| | | | | Add New + Bottom Left | | | | | | |
| | | | | Wait til complete | | | | | | |
| | | | Cerate Experiment | | | | | | | |
| | | | | Blank Experiment | | | | | | |
| | | | | | Saved Datasets->My Datasets -> | | | | | |
| | | | | | | Dataset | | | | |
| | | | | | | ** Important Note. In Python and Azure ML Studio, pieces are not available, just because they are there. You must run to that section, in order to view that section. | | | | |
| | | | | | Data Transformation -> Manipulation | | | | | |
| | | | | | | Clean Missing Data  -  Bad Data Matters... a lot.. and it is an art.. What do you want to drop... outliers.. zeros.... values that don't support your hypothesis | | | | |
| | | | | | | Edit Metadata  (Only if running ahead.. try changing quality to categorical) | | | | |
| | | | | | | Visualize Data ->  check missing values.. change to box&whiskers plot | | | | |
| | | | | | Select Columns in DataSet | | | | | |
| | | | | | | Launch Column Selector | | | | |
| | | | | | | Start Typing and Add | | | | |
| | | | | | | | LIVEAREA, FINBSMTNT, BASEMENT, CONDITION_ADJ, NOCARS, NUMBBDRM, NUMBBATHS, NETPRICE | | | |
| | | | | | Data Transformmation -> Sample and Split     - We always create 2 datasets... 1 for Training 1 for Testing | | | | | |
| | | | | | | Split Data | change Franction to .7 | | | |
| | | | | | | | Run then Visualize Data | | | |
| | | | | | Machine Learning ->  Initialize Model Regression | | | | | |
| | | | | | | Pick Algorithm, Poisson Regression or Boosted Decision Tree, off to left | | | | |
| | | | | | | | Hyperparameters | | | |
| | | | | | | Train Model | Connect Algorithm and first split of data | | | |
| | | | | | | | Add Column NETPRICE | | | |
| | | | | | Machine Learning -> Score | | | | | |
| | | | | | | Score Model | Connect train Model and split data | | | |
| | | | | | Machine Learning -> Evaluate | | | | | |
| | | | | | | Evaluate Model | | | | |
| | | | | Run | | | | | | |
| | | | | | 30ish seconds | | | | | |
| | | | | Evaluate | | | | | | |
| | | | | | View Output log | | | | | |
| | | | | | Grab JSON (Start at {"InputParameters ....  Duration":000.XX"} | | | | | |
| | | | | | Copy -> JSON HU | | | | | |
| | | | | | Coefficient of Determination | | | | | |
| | | | | | | | BDT : .33 | | | |
| | | | | | | | LR: .48 | LR: .38 w/ Categorical Quality ? | | |
| | | | | | | | PR .28 | | | |
| | | | | | | | NN -.05 | | | |
| | If Time - discuss | | | | | | | | | |
| | | | Setup Web Service | | | | | | | |
| | | | Improvements | | | | | | | |
| | | | | Better Data Clean | | | | | | |

| | | | | Clean Missing Data | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | |
| | | | Hyperparameters | | | | | | | | | | | |
| | | | more data | | | | | | | | | | | |
| | | | | Features from existing data | | | | | | | | | | |
| | | | | more rows - grab more neighborhoods, but pull out condos | | | | | | | | | | |
| | | | | more columns/features - lot size, elementary school rating | | | | | | | | | | |
| | | | | | | | | | | | | | | |
| | | If running ahead - https://www.anaconda.com/distribution/ 3.7 | | | | | | | | | | | | |