

A Mathematical Theory of Communication

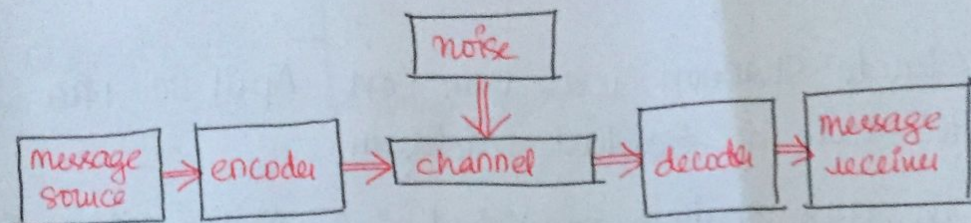
(1)

- Claude Shannon was born on April 30, 1916 in the town of Gaylord, Michigan.
- In 1948 Shannon published "A Mathematical Theory of Communication" which built on the foundations of other researchers at Bell Labs such as Henry Nyquist and R.V.L. Hartley.
- Shannon concentrated on 2 key questions:-
 - (i) determining the most efficient encoding of a message using a given alphabet in a noiseless environment.
 - (ii) understanding what additional steps need to be taken in the presence of noise.
- Shannon successfully solved this problem for a very abstract model of commⁿ systems that included both:-
 - (a) discrete (digital)
 - (b) continuous (analog) systems
- He developed a measure of the efficiency of a communication system called entropy.

ENTROPY — measures the amount of disorder in physical systems

↳ computed on the basis of mathematical statistical properties of message.

Shannon's communication model



→ His model had "applications not only in communication theory, but also in the theory of computing machines, the design of telephone exchanges and other fields".

→ each component can be treated independently as distinct mathematical models.

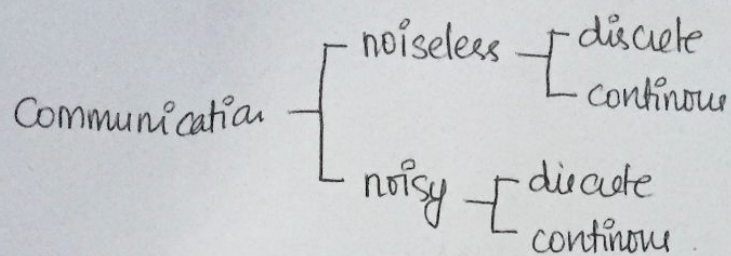
- Shannon's theory deals primarily with the encoder, channel, noise source and decoder.

Two fundamental ways to transmit message.

Discrete Signals
(represent only a finite number of different, recognizable states).

eg:- letters of the English alphabet.

Continuous Signals
(commonly used to transmit quantities that can vary over an infinite set of values)
eg:- sound.



CHANNEL - is the medium that carries the message.

Capacity of a discrete channel,

$$C = \lim_{T \rightarrow \infty} \frac{\log N(T)}{T}$$

where $N(T)$ - no: of allowed signals of duration T .

Entropy of an information source

Based on the probability mass function of each source symbol to be communicated, the Shannon entropy H in units of bits is given by,

$$H = - \sum_i P_i \log_2(P_i)$$

where P_i - is the probability of occurrence of the i -th possible value of the source symbol.

Problem 2: Scraping, Entropy and ICML papers.

ICML is a top research conference in Machine learning. Scrape all the pdfs of all ICML 2017 papers from <http://proceedings.mlr.press/v70/> (<http://proceedings.mlr.press/v70/>).

```
In [65]: # Import webdriver from selenium
from selenium import webdriver
import io
import requests
import re
import urllib
import wget
import textract
import os
#from collections import OrderedDict
from operator import itemgetter
#import random
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import enchant
import math
from selenium import webdriver
```

```
In [67]: chrome_path = r'/Users/aparnaaidith/Desktop/UGCA/Text-Analytics/chromedriver'
# prepend it with r because of the backslash characters
driver = webdriver.Chrome(chrome_path) # This will open up a chrome window
# Take a website that you want to scrape
driver.get("http://proceedings.mlr.press/v70/")
```

```
In [68]: linkElements= driver.find_elements_by_partial_link_text('Download PDF')
downloadLinks = []
for i in linkElements:
    downloadLinks.append(i.get_attribute('href'))
```



```
In [69]: count = 1
for link in downloadLinks:
    wget.download(link)
    print("Downloading PDF ... "+str(count) +' of '+str(len(downloadLinks))
    count = count + 1
```

```
Downloading PDF ... 1 of 434
Downloading PDF ... 2 of 434
Downloading PDF ... 3 of 434
Downloading PDF ... 4 of 434
Downloading PDF ... 5 of 434
Downloading PDF ... 6 of 434
Downloading PDF ... 7 of 434
Downloading PDF ... 8 of 434
Downloading PDF ... 9 of 434
Downloading PDF ... 10 of 434
Downloading PDF ... 11 of 434
Downloading PDF ... 12 of 434
Downloading PDF ... 13 of 434
Downloading PDF ... 14 of 434
Downloading PDF ... 15 of 434
Downloading PDF ... 16 of 434
Downloading PDF ... 17 of 434
Downloading PDF ... 18 of 434
Downloading PDF ... 19 of 434
```

```
In [70]: import os
path_to_pdfs = []
for file in os.listdir(os.getcwd()):
    if file.endswith(".pdf"):
        path_to_pdfs.append(file)
```

```
In [71]: from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
from pdfminer.converter import TextConverter
from pdfminer.layout import LAParams
from pdfminer.pdfpage import PDFPage

def convert_pdf_to_txt(path):
    rsrcmgr = PDFResourceManager()
    retstr = io.StringIO()
    codec = 'utf-8'
    laparams = LAParams()
    device = TextConverter(rsrcmgr, retstr, codec=codec, laparams=laparams)
    fp = open(path, 'rb')
    interpreter = PDFPageInterpreter(rsrcmgr, device)
    password = ""
    maxpages = 0
    caching = True
    pagenos = set()

    for page in PDFPage.get_pages(fp, pagenos, maxpages=maxpages,
                                  password=password,
                                  caching=caching,
                                  check_extractable=True):
        interpreter.process_page(page)

    text = retstr.getvalue()

    fp.close()
    device.close()
    retstr.close()
    return text

#text = convert_pdf_to_txt('achab17a.pdf')
```

```
In [79]: textTest = convert_pdf_to_txt('ruggieri17a.pdf')
textTest
```

```
-----
--
PSTypeError                                Traceback (most recent call las
t)
<ipython-input-79-8aa2a89baec8> in <module>()
----> 1 textTest = convert_pdf_to_txt('ruggieri17a.pdf')
      2 textTest

<ipython-input-71-735bb2b3edd4> in convert_pdf_to_txt(path)
      21             caching=caching,
      22             check_extractable=True):
----> 23         interpreter.process_page(page)
      24
      25         text = retstr.getvalue()

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in process_p
age(self, page)
      832             ctm = (1, 0, 0, 1, -x0, -y0)
      833             self.device.begin_page(page, ctm)
--> 834             self.render_contents(page.resources, page.contents, ctm=c
tm)
      835             self.device.end_page(page)
      836             return

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in render_co
ntents(self, resources, streams, ctm)
      844             self.init_resources(resources)
      845             self.init_state(ctm)
--> 846             self.execute(list_value(streams))
      847             return
      848

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in execute(s
elf, streams)
      868             logging.debug('exec: %s %r', name, args)
      869             if len(args) == nargs:
--> 870                 func(*args)
      871             else:
      872                 logging.debug('exec: %s', name)

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in do_Do(sel
f, xobjid)
      809             resources = dict_value(xobjres) if xobjres else self.
resources.copy()
      810             self.device.begin_figure(xobjid, bbox, matrix)
--> 811             interpreter.render_contents(resources, [xobj], ctm=mu
lt_matrix(matrix, self.ctm))
      812             self.device.end_figure(xobjid)
      813             elif subtype is LITERAL_IMAGE and 'Width' in xobj and 'He
ight' in xobj:

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in render_co
ntents(self, resources, streams, ctm)
      842             logging.info('render_contents: resources=%r, streams=%r,
```

```

ctm=%r',
843             resources, streams, ctm)
--> 844         self.init_resources(resources)
845         self.init_state(ctm)
846         self.execute(list_value(streams))

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in init_resources(self, resources)
348             objid = spec.objid
349             spec = dict_value(spec)
--> 350             self.fontmap[fontid] = self.rsrcmgr.get_font(
objid, spec)
351         elif k == 'ColorSpace':
352             for (csid, spec) in six.iteritems(dict_value(v)):

/anaconda3/lib/python3.6/site-packages/pdfminer/pdfinterp.py in get_font(self, objid, spec)
180         if subtype in ('Type1', 'MMType1'):
181             # Type1 Font
--> 182             font = PDFType1Font(self, spec)
183         elif subtype == 'TrueType':
184             # TrueType Font

/anaconda3/lib/python3.6/site-packages/pdfminer/pdffont.py in __init__(self, rsrcmgr, spec)
573     def __init__(self, rsrcmgr, spec):
574         try:
--> 575             self.basefont = literal_name(spec['BaseFont'])
576         except KeyError:
577             if STRICT:

/anaconda3/lib/python3.6/site-packages/pdfminer/psparser.py in literal_name(x)
136         if not isinstance(x, PSLiteral):
137             if STRICT:
--> 138                 raise PSTypeError('Literal required: %r' % x)
139             else:
140                 name=x

```

PSTypeError: Literal required: <PDFObjRef:239>


```
In [72]: path_to_pdfs
         'suggala17a.pdf',
         'masegosa17a.pdf',
         'ubaru17a.pdf',
         'bernstein17a.pdf',
         'heckel17a.pdf',
         'pakman17a.pdf',
         'tsakiris17a.pdf',
         'begon17a.pdf',
         'budden17a.pdf',
         'jabbari17a.pdf',
         'chen17f.pdf',
         'jin17a.pdf',
         'macglashan17a.pdf',
         'balduzzi17b.pdf',
         'kohler17a.pdf',
         'hu17b.pdf',
         'tu17a.pdf',
         'tan17a.pdf',
         'chen17d.pdf',
         'jing17a.pdf',
         'ruggieri17a.pdf'
```

```
In [80]: text = ''
         count = 0
         for pdf in path_to_pdfs:
             if count < 10 and pdf != 'ruggieri17a.pdf':
                 print("Parsing... "+str(count)+' of 10 - '+pdf)
                 text = text + convert_pdf_to_txt(pdf)
                 count = count + 1
         print("Done!")
```

```
Parsing... 0 of 10 - suggala17a.pdf
Parsing... 1 of 10 - miller17a.pdf
Parsing... 2 of 10 - ali17a.pdf
Parsing... 3 of 10 - munkhdalail7a.pdf
Parsing... 4 of 10 - chen17a.pdf
Parsing... 5 of 10 - anschell17a.pdf
Parsing... 6 of 10 - khasanova17a.pdf
Parsing... 7 of 10 - sugiyama17a.pdf
Parsing... 8 of 10 - winner17a.pdf
Parsing... 9 of 10 - chowdhury17a.pdf
Done!
```

```
In [81]: text
Computing, KAIST, Daejeon, South Korea (SAILLIES, Seoul, South Korea. (no
correspondence to: Arun Sai Suggala <asuggala@andrew.cmu.edu>, Eunho Yang
<eun-\\nhoy@kaist.ac.kr>.)
Proceedings of the 34 th International Confer-
ence on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 20
17 by the author(s).
variables. This conditional independence structu-
re pro-\\vides us with useful insights about how different variables\\nint-
eract with each other. As a result MRFs are extensively\\nused in a variet-
y of fields, including Natural Language Pro-\\ncessing (Manning & Sch\\utze,
1999), Biology (Friedman,\\n2004) and Medicine (Allen & Liu, 2012).\\nPopul-
ar parametric families in the general class of MRFs\\nare Gaussian graphic-
al models (Speed & Kiiveri, 1986;\\nRue & Held, 2005) for continuous (and
bell-shaped) data,\\nIsing and discrete graphical models (Ising, 1925; Jal-
ali\\net al., 2011) for nominal data, and mixed cases of these two\\ninstan-
ces (Lauritzen & Wermuth, 1989; Yang et al., 2014).\\nHowever, variables t-
hat occur in many real world appli-\\ncations have ordered categorical sca-
les. For example, in\\nmedical data, diseases are graded from mild to fata-
l, sever-\\nity of an injury is rated from mild injury to death, stages of
\\na disease is rated from I to III. Ordinal variables also occur\\nvery co-
mmonly in data collected from surveys. For exam-\\nple, each subject takin-
g a survey could be asked to respond\\nto a question using categories such
```

```
In [82]: #text_clean = re.sub(r'\\d+', ' ', str(text))
text_clean1 = re.sub('[^a-zA-Z-]+', ' ', str(text))
text_clean = re.sub("\\s*-\\s*\\n", "", str(text_clean1))
```

```
In [83]: text_clean
```

```
Out[83]: 'Ordinal Graphical Models A Tale of Two Approaches Arun Sai Suggala Eunho
Yang Pradeep Ravikumar Abstract Undirected graphical models or Markov ran-
dom elds MRFs are widely used for modeling mul- tivariate probability dis-
tributions Much of the work on MRFs has focused on continuous vari- ables
and nominal variables that is unordered categorical variables However dat-
a from many real world applications involve ordered categor- ical variabl-
es also known as ordinal variables e g movie ratings on Net ix which can
be or- dered from to stars With respect to uni- variate ordinal distribut-
ions as we detail in the paper there are two main categories of distri- b-
utions while there have been efforts to extend these to multivariate ordi-
nal distributions the re- sulting distributions are typically very comple-
x with either a large number of parameters or with non-convex likelihoods
While there have been some work on tractable approximations these do not
come with strong statistical guarantees and moreover are relatively compu-
tationally expen- sive In this paper we theoretically investigate two cla-
sses of graphical models for ordinal data corresponding to the two main c-
ategories of uni- variate ordinal distributions In contrast to pre- vious
work our theoretical developments allow us to provide correspondingly two
classes of esti- mators that are not only computationally ef cient but al-
```

```
In [84]: from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
tokens = word_tokenize(text_clean)
tokens
```

```
Out[84]: ['Ordinal',
'Graphical',
'Models',
'A',
'Tale',
'of',
'Two',
'Approaches',
'Arun',
'Sai',
'Suggala',
'Eunho',
'Yang',
'Pradeep',
'Ravikumar',
'Abstract',
'Undirected',
'graphical',
'models',
']
```

```
In [ ]: text_clean = re.sub(r'\d+', ' ', str(text))
```

```
In [85]: punctuations = ['(', ')', ';', ':', '[', ']', ',', '.', '&', '\n', '=', '€']
stop_words = stopwords.words('english')
englishDictionary = enchant.Dict("en_US")
keywords = [word for word in tokens if not word in stop_words and not word
```

```
In [86]: keywords
```

```
Out[86]: ['Ordinal',
'Graphical',
'Models',
'A',
'Tale',
'Two',
'Approaches',
'Yang',
'Abstract',
'Undirected',
'graphical',
'models',
'Markov',
'random',
'widely',
'used',
'modeling',
'probability',
'distributions',
']
```

```
In [16]: ' ' in keywords
```

```
Out[16]: False
```

```
In [87]: # for i in range(len(keywords)):
# #     if(len(keywords[i]) >= 1) :
#         j = 0
#         print(keywords[i])
# #         if(keywords[i][-1] == '-'):
# #             j+=1
# #             keywords[i] = keywords[i][: -1] + keywords[i]
# #             del(keywords[i+1])
# print(j)

# f = keywords[:30]
# f
j = 0
for i in range(len(keywords)):

    if(i < len(keywords)):
        if(keywords[i][-1] == '-'):
            keywords[i] = keywords[i][: -1] + keywords[i]
            del(keywords[i+1])
```

```
In [88]: for i in range(len(keywords)):
        if(i < len(keywords)):
            if(keywords[i][-1] == '-'):
                del(keywords[i])
```

```
In [89]: keywords
        'detail',
        'paper',
        'two',
        'main',
        'categories',
        'efforts',
        'extend',
        'multivariate',
        'ordinal',
        'distributions',
        'distributions',
        'typically',
        'complex',
        'either',
        'large',
        'number',
        'parameters',
        'likelihoods',
        'While',
        'work',
        .....
```



```
In [92]: counts = dict()
englishDictionary = enchant.Dict("en_US")

for word in keywords:
    #remove one letter words, stopwords, and make sure that the word is act
    if len(word) > 1 and word not in ['The', 'In', 'et', 'In', 'We']:
        if word in counts:
            counts[word] += 1
        else:
            counts[word] = 1

sortedWordsList = sorted(counts.items(), key = itemgetter(1), reverse = True)
count = 1
for word in sortedWordsList:
    if count < 11:
        print(word)
    count = count + 1

('model', 213)
('learning', 140)
('set', 137)
('models', 131)
('function', 128)
('algorithm', 124)
('distribution', 114)
('data', 106)
('Learning', 104)
('method', 104)
```

Problem 2.2

Let Z be a randomly selected word in a randomly selected ICML paper. Estimate the entropy of Z .

```
In [95]: englishDictionary = enchant.Dict("en_US")
random_pdf = random.choice(path_to_pdfs)
text = convert_pdf_to_txt(random_pdf)
```

```
In [ ]: text_clean1 = re.sub('[^a-zA-Z-]+', ' ', str(text))
text_clean = re.sub("\s*-\s*\n", "", str(text_clean1))
```

```
In [ ]: from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
tokens = word_tokenize(text_clean)
tokens
```

```
In [ ]: punctuations = ['(', ')', ';', ':', '[', ']', ',', '.', '&', '\n', '=', '€']
stop_words = stopwords.words('english')
englishDictionary = enchant.Dict("en_US")
keywords = [word for word in tokens if not word in stop_words and not word i
```

```
In [ ]: for i in range(len(keywords)):
        if(i < len(keywords)):
            if(keywords[i][-1] == '-'):
                keywords[i] = keywords[i][:-1] + keywords[i]
            del(keywords[i+1])
```

```
In [ ]: for i in range(len(keywords)):
        if(i < len(keywords)):
            if(keywords[i][-1] == '-'):
                del(keywords[i])
```

```
In [ ]: counts = dict()
englishDictionary = enchant.Dict("en_US")

for word in keywords:
    #remove one letter words, stopwords, and make sure that the word is actu
    if len(word) > 1 and word not in ['The', 'In', 'et', 'In', 'We']:
        if word in counts:
            counts[word] += 1
        else:
            counts[word] = 1

sortedWordsList = sorted(counts.items(), key = itemgetter(1), reverse = True)
count = 1
for word in sortedWordsList:
    if count < 11:
        print(word)
    count = count + 1
```

```
In [113]: randomWord = random.choice(keywords)
```

```
In [114]: randomWord
```

```
Out[114]: 'converge'
```

```
In [115]: wordFreqDict = {}
for word in keywords:
    if word in wordFreqDict:
        wordFreqDict[word] += 1
    else:
        wordFreqDict[word] = 1

#Get random word's frequency
randomWordFrequency = wordFreqDict[randomWord]
#print(wordFreqDict) # this prints out words
```

```
In [116]: randomWordFrequency
```

```
Out[116]: 12
```

```
In [118]: totalWordCount = len(keywords)
totalWordCount
```

```
Out[118]: 34446
```

```
In [119]: wordFreqDict2 = dict(wordFreqDict)
wordFreqDict2
```

```
Out[119]: {'Ordinal': 13,
'Graphical': 20,
'Models': 42,
'A': 267,
'Tale': 10,
'Two': 12,
'Approaches': 10,
'Yang': 17,
'Abstract': 10,
'Undirected': 2,
'graphical': 33,
'models': 131,
'Markov': 18,
'random': 75,
'widely': 5,
'used': 68,
'modeling': 12,
'probability': 61,
'distributions': 92,
'words': 1}
```

```
In [120]: entropySum = 0

for word in keywords:
    #pdb.set_trace()
    wordFrequency = wordFreqDict[word]
    entropy = -(wordFrequency/totalWordCount)*math.log(wordFrequency/total
    entropySum = entropySum + entropy

print(entropySum)
```

```
371.3629104147223
```

Question 1 & 2 :

```
In [4]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib

import matplotlib.pyplot as plt
from scipy.stats import skew
from scipy.stats.stats import pearsonr

%config InlineBackend.figure_format = 'retina' #set 'png' here when work
ing on notebook
%matplotlib inline
```

```
In [5]: train = pd.read_csv("/Users/aparnaaidith/Documents/FALL 2018/Business Da
ta Science/all/train.csv")
test = pd.read_csv("/Users/aparnaaidith/Documents/FALL 2018/Business Dat
a Science/all/test.csv")
```

```
In [6]: train.head()
```

Out[6]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Util
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	Al
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	Al
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	Al
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	Al
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	Al

5 rows × 81 columns

```
In [7]: all_data = pd.concat((train.loc[:, 'MSSubClass': 'SaleCondition'],
                             test.loc[:, 'MSSubClass': 'SaleCondition']))
```

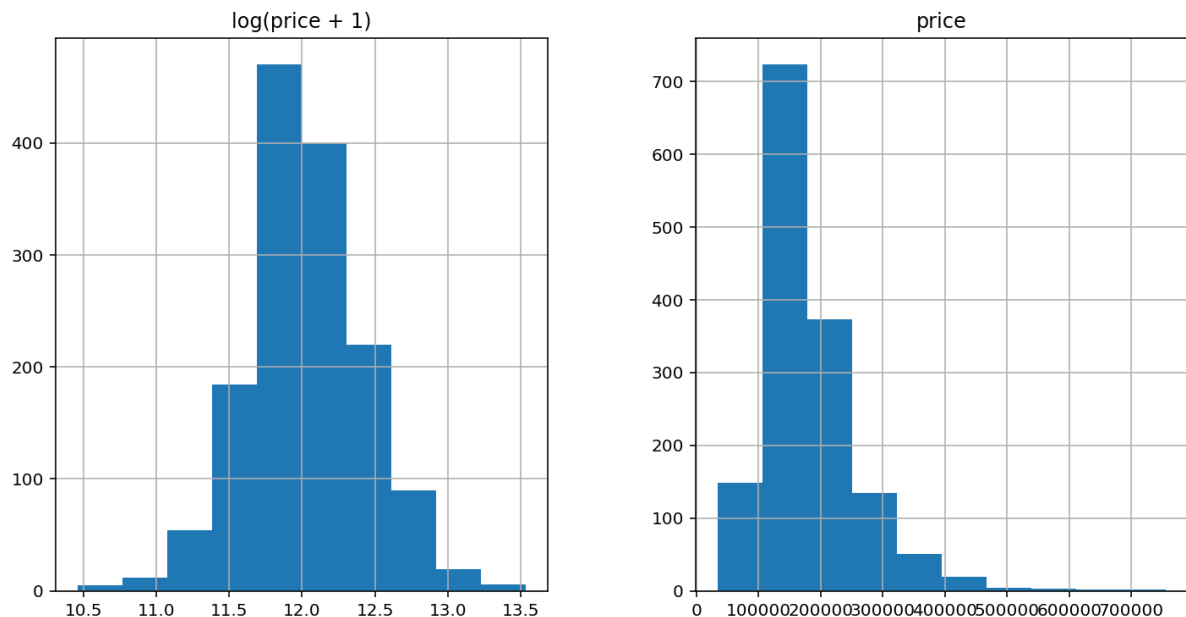


```
In [8]: import matplotlib

import matplotlib.pyplot as plt

matplotlib.rcParams['figure.figsize'] = (12.0, 6.0)
prices = pd.DataFrame({"price":train["SalePrice"], "log(price + 1)":np.log1p(train["SalePrice"])})
prices.hist()
```

```
Out[8]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x105f6df98>,
               <matplotlib.axes._subplots.AxesSubplot object at 0x111fa5518
               >]],
              dtype=object)
```



```
In [9]: #log transform the target:
train["SalePrice"] = np.log1p(train["SalePrice"])

#log transform skewed numeric features:
numeric_feats = all_data.dtypes[all_data.dtypes != "object"].index

skewed_feats = train[numeric_feats].apply(lambda x: skew(x.dropna())) #compute skewness
skewed_feats = skewed_feats[skewed_feats > 0.75]
skewed_feats = skewed_feats.index

all_data[skewed_feats] = np.log1p(all_data[skewed_feats])
```

```
In [10]: all_data = pd.get_dummies(all_data)
```

```
In [11]: #filling NA's with the mean of the column:
all_data = all_data.fillna(all_data.mean())
```

```
In [12]: #creating matrices for sklearn:
X_train = all_data[:train.shape[0]]
X_test = all_data[train.shape[0]:]
y = train.SalePrice

X_train.shape
```

```
Out[12]: (1460, 288)
```

```
In [13]: X_test.shape
```

```
Out[13]: (1459, 288)
```

```
In [14]: from sklearn.linear_model import Ridge, RidgeCV, ElasticNet, LassoCV, La
ssoLarsCV
from sklearn.model_selection import cross_val_score

from sklearn.metrics import mean_squared_error

model_ridge = Ridge(alpha=0.1)
model_ridge.fit(X_train, y)

# def rmse_cv(model):

#     rmse = np.sqrt(mean_squared_error(X_test, y))
#     return(rmse)
```

```
Out[14]: Ridge(alpha=0.1, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=None, solver='auto', tol=0.001)
```

```
In [15]: y_hat_train = model_ridge.predict(X_train)
new_y_hat_train = np.expml(y_hat_train)
```

```
In [16]: rmse_train = np.sqrt(mean_squared_error(y, y_hat_train))
```

```
In [17]: rmse_train
```

```
Out[17]: 0.09211955585640517
```

```
In [18]: y_hat_test = model_ridge.predict(X_test)
new_y_hat_test = np.expml(y_hat_test)

rmse_test = np.sqrt(mean_squared_error(y, y_hat_train))
rmse_test
```

```
Out[18]: 0.09211955585640517
```

```
In [19]: solution = pd.DataFrame({"Id":(test.Id) , "SalePrice": new_y_hat_test})
solution.to_csv("/Users/aparnaaidith/Documents/FALL 2018/Business Data S
cience/all/ridge_sol.csv", index = False)
```

Question 3 :

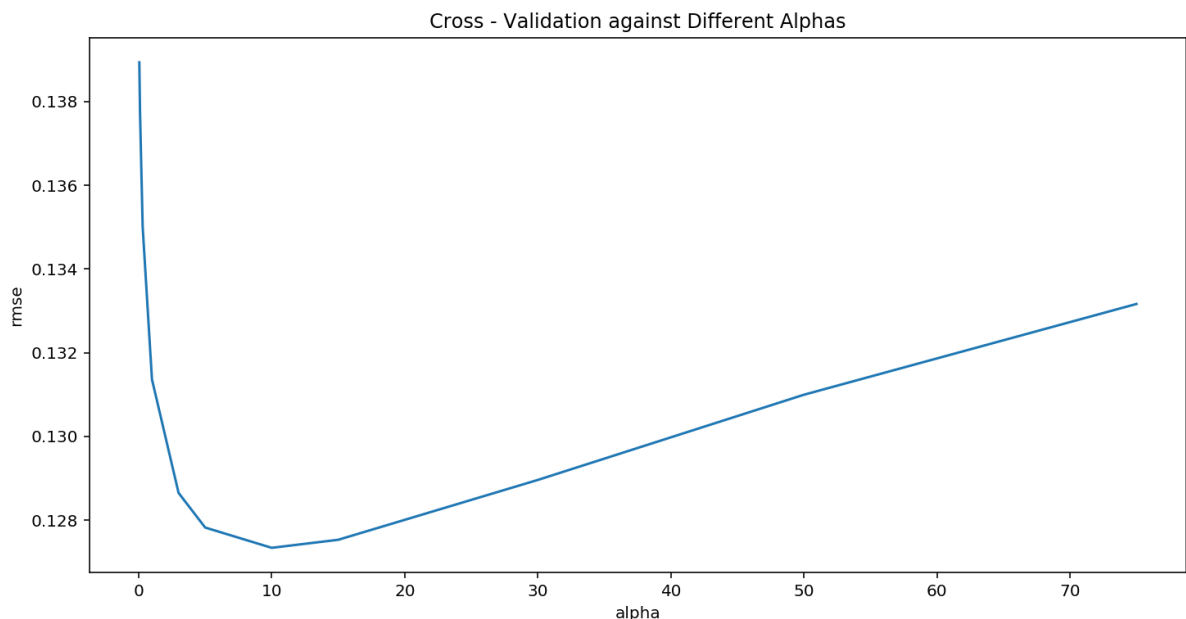
```
In [20]: from sklearn.linear_model import Ridge, RidgeCV, ElasticNet, LassoCV, LassoLarsCV, Lasso
from sklearn.model_selection import cross_val_score

def rmse_cv(model):
    rmse= np.sqrt(-cross_val_score(model, X_train, y, scoring="neg_mean_squared_error", cv = 5))
    return(rmse)
```

```
In [21]: alphas = [0.05, 0.1, 0.3, 1, 3, 5, 10, 15, 30, 50, 75]
cv_ridge = [rmse_cv(Ridge(alpha = alpha)).mean()
             for alpha in alphas]
```

```
In [22]: cv_ridge = pd.Series(cv_ridge, index = alphas)
cv_ridge.plot(title = "Cross - Validation against Different Alphas")
plt.xlabel("alpha")
plt.ylabel("rmse")
```

Out[22]: Text(0, 0.5, 'rmse')



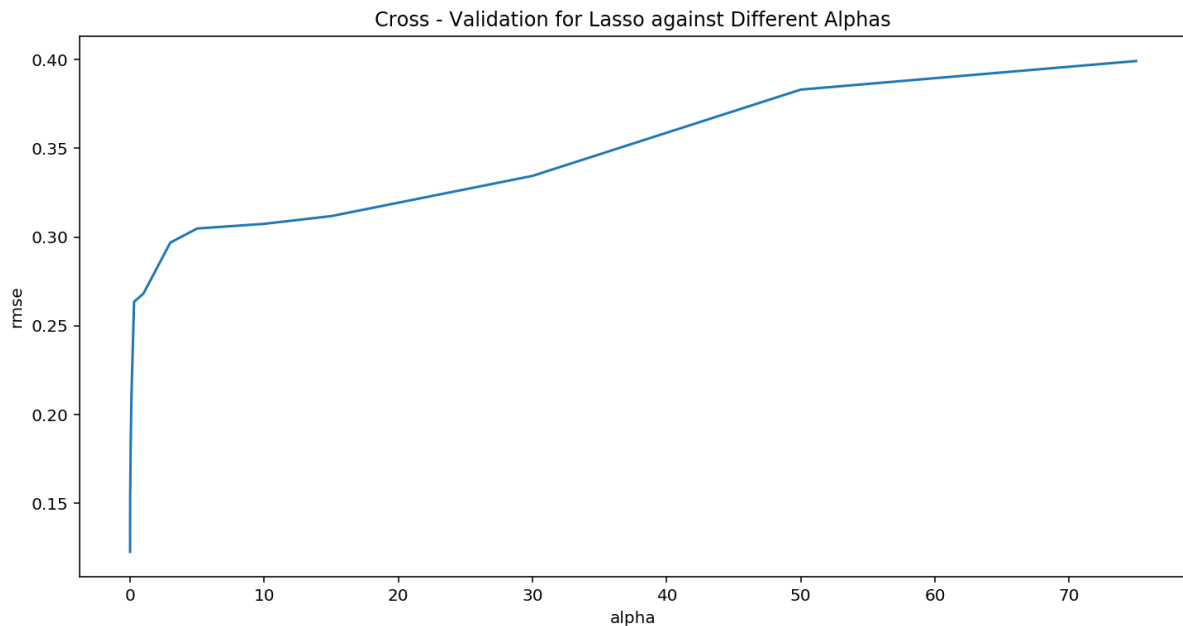
```
In [23]: cv_ridge.min()
```

Out[23]: 0.12733734668670763

```
In [24]: #Tuning: lets try various Alphas and do cross validation to see which works better

alphas = [.0005, .001, .005, .01, 0.05, 0.1, 0.3, 1, 3, 5, 10, 15, 30, 50, 75]
cv_lasso = [rmse_cv(Lasso(alpha = alpha)).mean()
             for alpha in alphas]
cv_lasso = pd.Series(cv_lasso, index = alphas)
cv_lasso.plot(title = "Cross - Validation for Lasso against Different Alphas")
plt.xlabel("alpha")
plt.ylabel("rmse")
```

Out[24]: Text(0, 0.5, 'rmse')



```
In [66]: cv_lasso.min()
```

Out[66]: 0.12256735885048149

```
In [76]: alphas = [0.05, 0.1, 0.3, 1, 3, 5, 10, 15, 30, 50, 75]
cv_ridge = [[alpha, rmse_cv(Ridge(alpha=alpha)).mean()]
             for alpha in alphas]
cv_lasso = [[alpha, rmse_cv(Lasso(alpha=alpha)).mean()]
             for alpha in alphas]

# find the alpha with the smallest score for ridge regression

print('Ridge alpha with minimum rmse:', min(cv_ridge, key=lambda x: x[1]))
print('Lasso alpha with minimum rmse:', min(cv_lasso, key=lambda x: x[1]))
```

```
Ridge alpha with minimum rmse: [10, 0.12733734668670763]
Lasso alpha with minimum rmse: [0.05, 0.1840376201101268]
```


Question 4 :

```
In [25]: model_lasso = LassoCV(alphas = [1, 0.1, 0.001, 0.0005]).fit(X_train, y)

/Users/aparnaaidith/py_37_env/lib/python3.7/site-packages/sklearn/model_selection/_split.py:1943: FutureWarning: You should specify a value for 'cv' instead of relying on the default value. The default value will change from 3 to 5 in version 0.22.
  warnings.warn(CV_WARNING, FutureWarning)
```

```
In [26]: coef = pd.Series(model_lasso.coef_, index = X_train.columns)
```

```
In [27]: print("Lasso picked " + str(sum(coef != 0)) + " variables and eliminated
the other " + str(sum(coef == 0)) + " variables")
```

Lasso picked 111 variables and eliminated the other 177 variables

```
In [28]: imp_coef = pd.concat([coef.sort_values().head(10),
                             coef.sort_values().tail(10)])
```

```
In [29]: matplotlib.rcParams['figure.figsize'] = (8.0, 10.0)
imp_coef.plot(kind = "barh")
plt.title("Coefficients in the Lasso Model")
```

```
Out[29]: Text(0.5, 1.0, 'Coefficients in the Lasso Model')
```



Question 6 :

```
In [79]: import xgboost as xgb
```

```
In [80]: from xgboost import XGBRegressor

dtrain = xgb.DMatrix(X_train, label = y)
dtest = xgb.DMatrix(X_test)

params = {"max_depth":2, "eta":0.1}
model = xgb.cv(params, dtrain, num_boost_round=500, early_stopping_rounds=100)
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

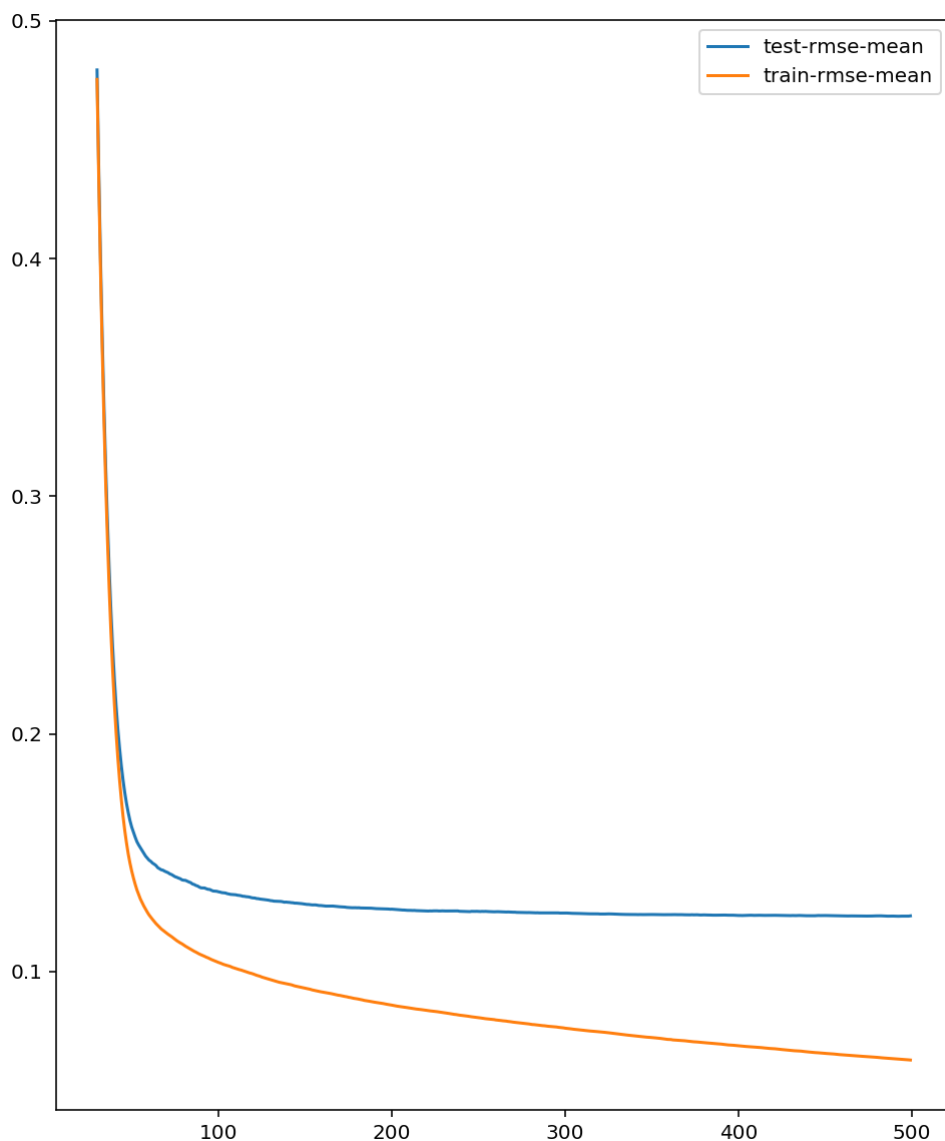
[illegible]

```
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 4 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
[10:18:28] src/tree/updater_prune.cc:74: tree pruning end, 1 roots, 6 e
xtra nodes, 0 pruned nodes, max_depth=2
```



```
In [81]: model.loc[30:,[ "test-rmse-mean", "train-rmse-mean"]].plot()
```

```
Out[81]: <matplotlib.axes._subplots.AxesSubplot at 0x115f80a20>
```



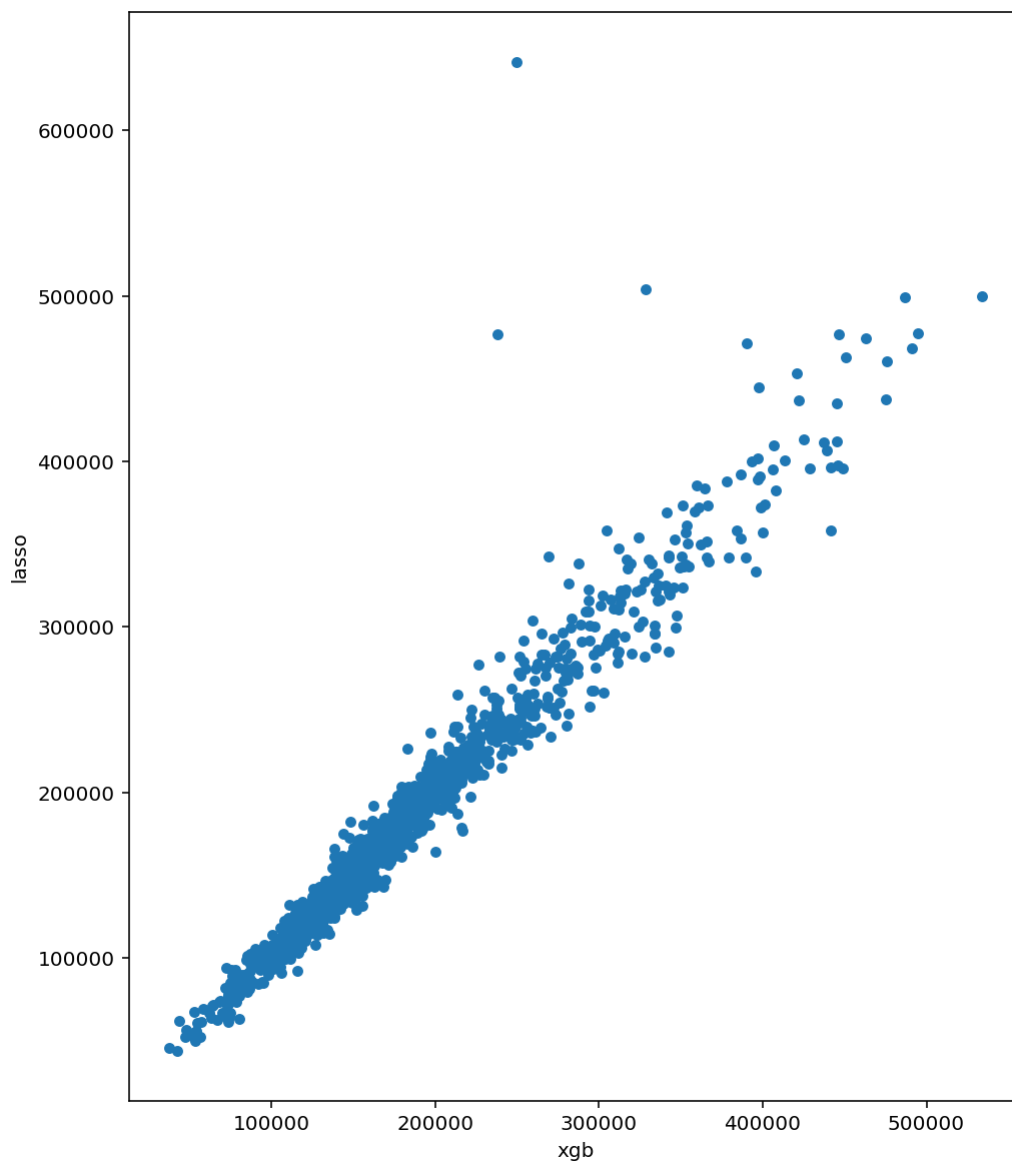
```
In [82]: model_xgb = xgb.XGBRegressor(n_estimators=360, max_depth=2, learning_rate=0.1) #the params were tuned using xgb.cv
model_xgb.fit(X_train, y)
```

```
Out[82]: XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=
0,
max_depth=2, min_child_weight=1, missing=None, n_estimators=360,
n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=True, subsample=1)
```

```
In [83]: xgb_preds = np.expml(model_xgb.predict(X_test))
lasso_preds = np.expml(model_lasso.predict(X_test))
```

```
In [84]: predictions = pd.DataFrame({"xgb":xgb_preds, "lasso":lasso_preds})  
predictions.plot(x = "xgb", y = "lasso", kind = "scatter")
```

```
Out[84]: <matplotlib.axes._subplots.AxesSubplot at 0x11613ca90>
```



Question 5 :

```
In [85]: from sklearn.linear_model import Lasso  
  
lasso_mod = Lasso(alpha=.1)  
lasso_mod.fit(X_train, y)  
y_pred_lasso = lasso_mod.predict(X_train)  
  
ridge_mod = Ridge(alpha=0.1)  
ridge_mod.fit(X_train, y)  
y_pred_ridge = ridge_mod.predict(X_train)
```

```
In [86]: X_train_new = X_train.copy()  
X_train_new['y_ridge'] = y_pred_ridge  
X_train_new['y_lasso'] = y_pred_lasso
```

```
In [87]: # create and train the model with the augmented data, then make a predic  
tion  
  
new_ridge_mod = Ridge(alpha=.1)  
new_ridge_mod.fit(X_train_new, y)  
y_hat_ens = new_ridge_mod.predict(X_train_new)  
# compute the rmse for our augmented data  
  
aug_rmse = np.sqrt(mean_squared_error(y, y_hat_ens))  
aug_rmse
```

```
Out[87]: 0.09185410763520462
```

```
In [ ]:
```