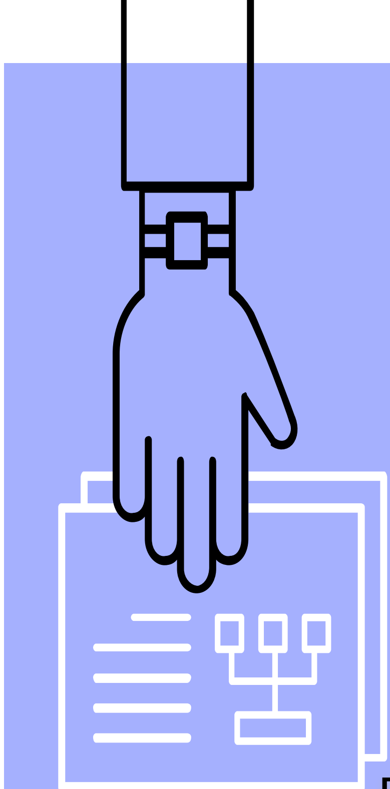
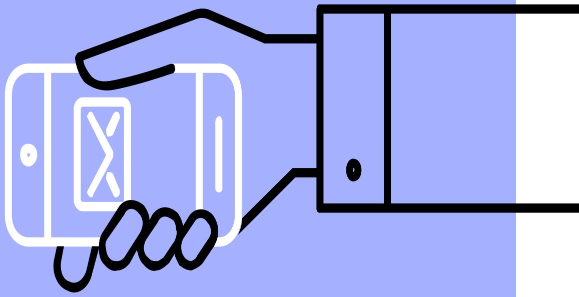
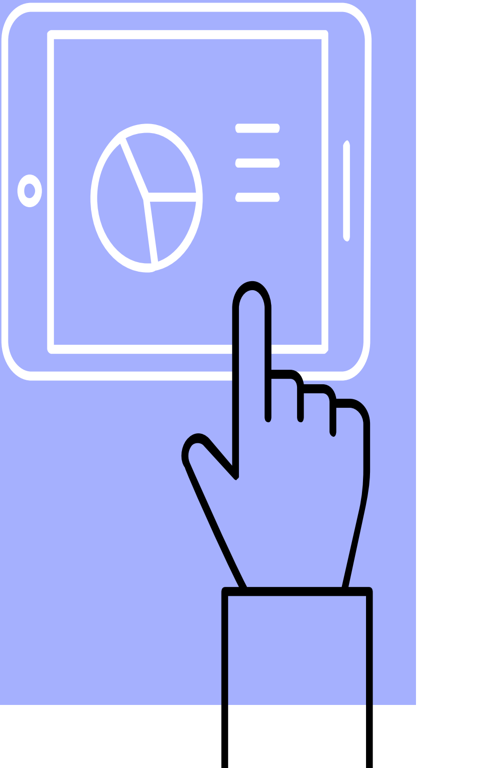
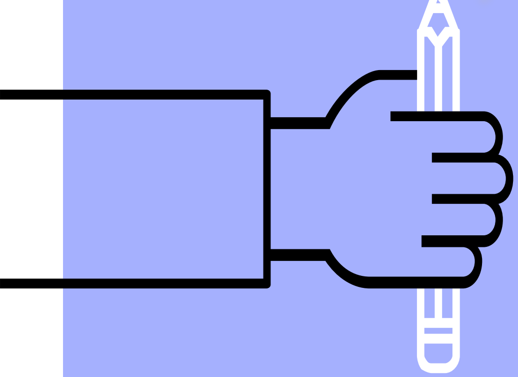




Multivariate Regression with Categorical Responses

Robert Nakano	robertnakano@gmail.com
Sunny Jelokhani	afsanehkhani35@gmail.com
Ziyi Wang	joannewangziyi@gmail.com





Github:
https://github.com/bobbyinfj/projects/tree/master/510%20Project-%20Categorical%20Response

Question?

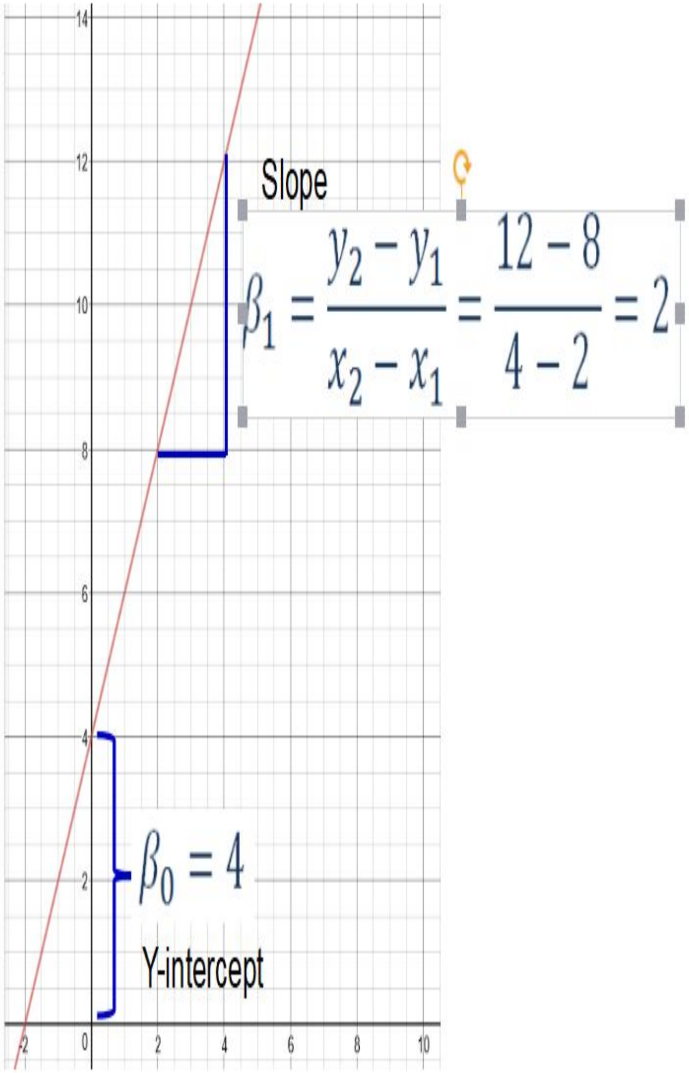
We've learned how to model with a numeric response variable

But, what if the response variable (Y) is not numeric?

Linear Equation Example

$$y = \beta_0 + \beta_1 x$$

$$y = 4 + 2x$$



Slope

$$\beta_1 = \frac{y_2 - y_1}{x_2 - x_1} = \frac{12 - 8}{4 - 2} = 2$$

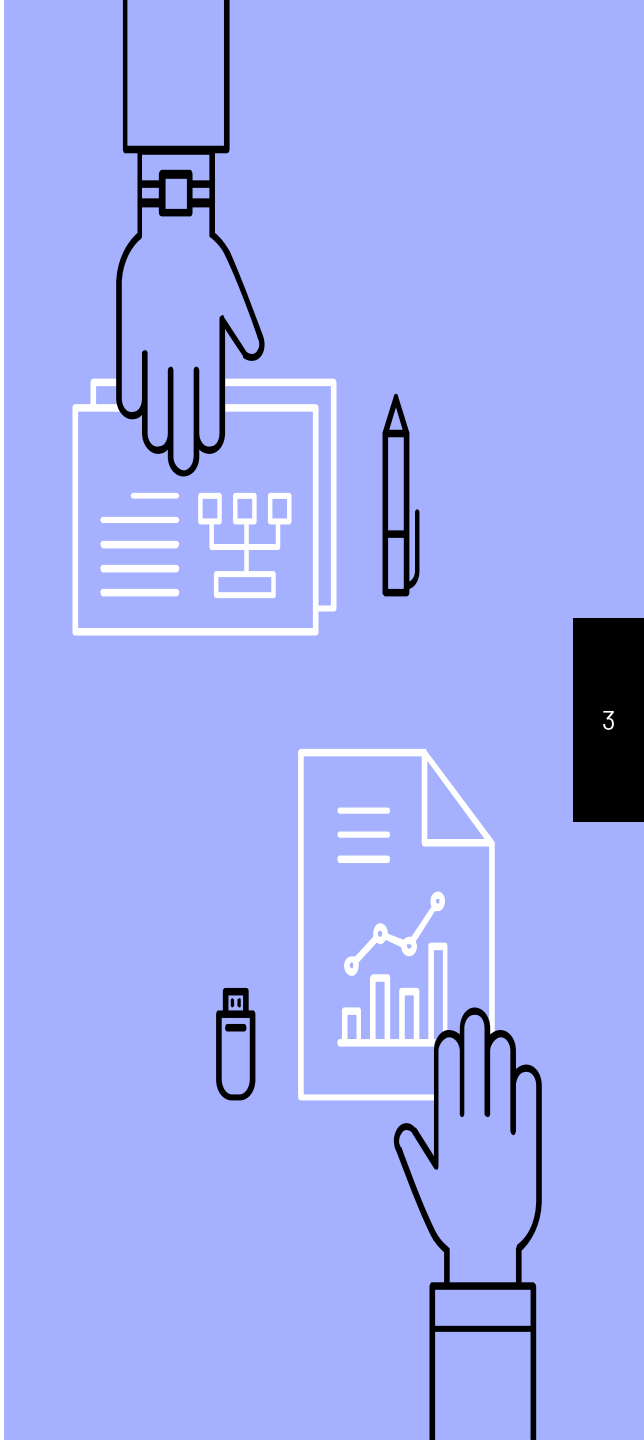
$$\beta_0 = 4$$

Y-intercept

Outline:

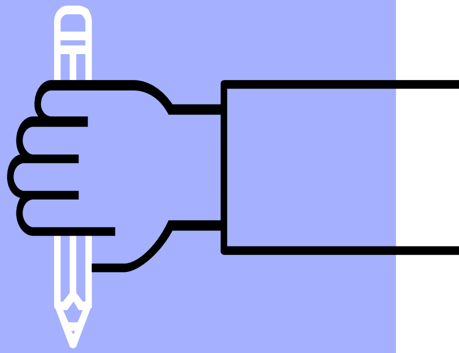
❑ **Overview of Categorical Responses**

❑ **Case Study : Predicting personalities (binary)**



Categorical Data & Reponses

What is a categorical response?

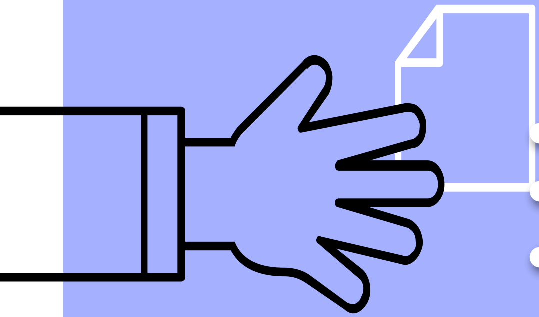


Response data that is measured by categories instead of continuously.

Also called: Qualitative (vs. Quantitative),

Qualitative:

- Color of a sample
- Texture of a surface
- Aroma of a reaction



Quantitative:

- Mass of a sample
- Length of a piece of wire
- Molecules in a mole

Types of Categorical Responses

Binary: 1 or 0

The response variable is one of two things.

- Smoker or Non-smoker
- Netflix Thumbs up or Thumbs down
- People who dance in front of the mirror or people who don’t dance in front of the mirror

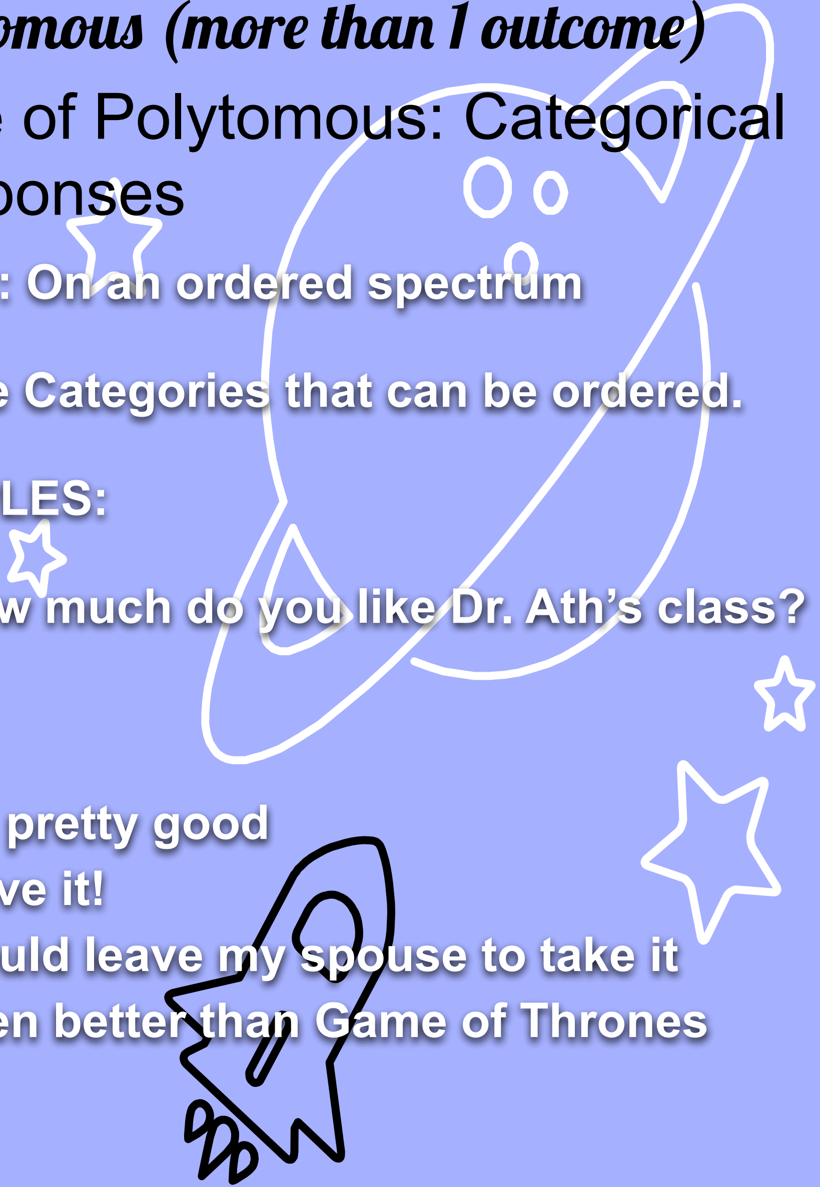
BIG

CONCEPT

- If a variable is not binary, it is Polytomous

Polytomous (more than 1 outcome)

- Type of Polytomous: Categorical Responses



Ordinal: On an ordered spectrum

Multiple Categories that can be ordered.

EXAMPLES:

E.g. How much do you like Dr. Ath's class?

(1 - 4)

1. **It's pretty good**
2. **I love it!**
3. **Would leave my spouse to take it**
4. **Even better than Game of Thrones**

Interval: numerical distance between data points

Types of Categorical Responses: Nominal

Nominal categorical responses have no order (unlike ordinal)

Worst Television Show?

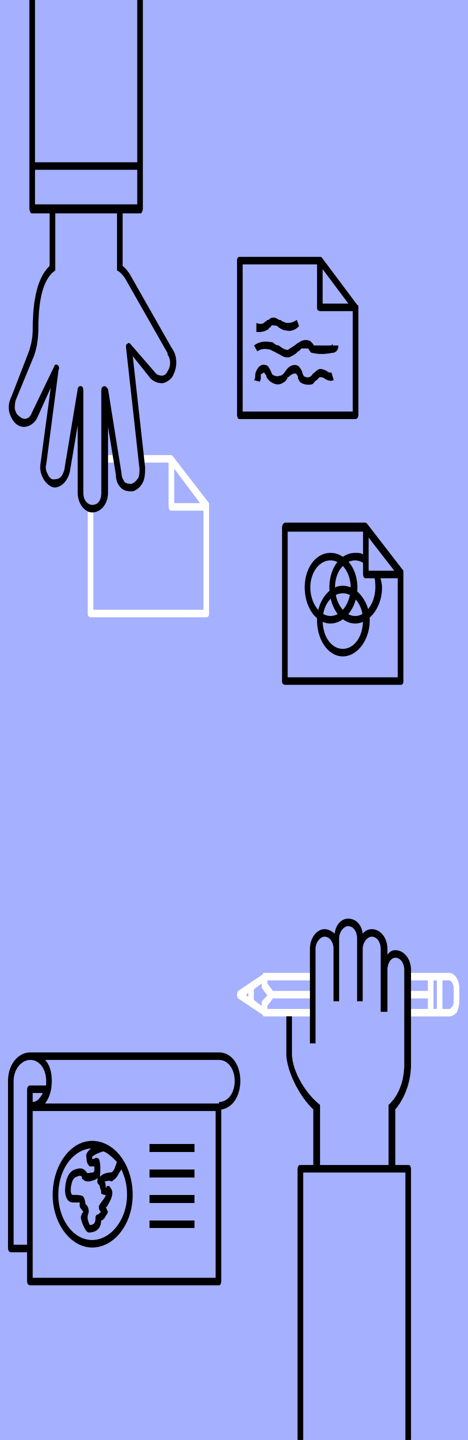
iZombie, Santa Clarita Diet, Ironfist

Who is your favorite singer?

Taylor Swift, The Biebs, Beyonce, Shakira, Robert Nakano

Favorite Dessert?

Tiramisu, Chocolate Souffle, Beef jerky, Froyo!



How do you predict a Categorical Response?

How Do You Predict a Categorical Response?

It depends on what kind:

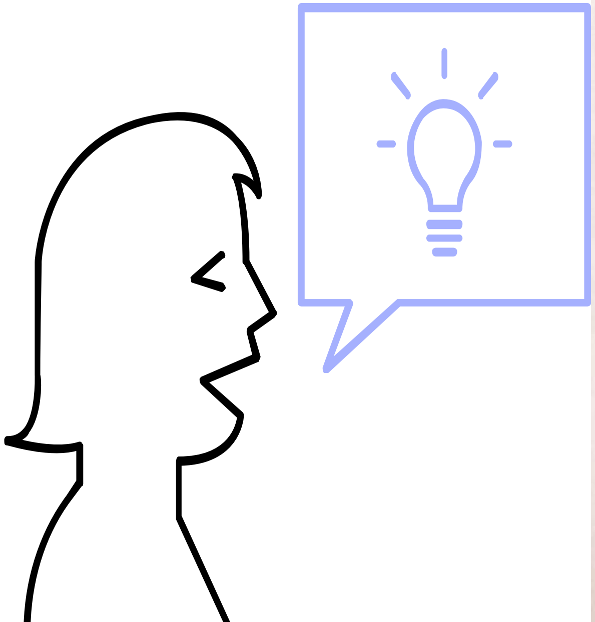
Binary-> Logistic Regression

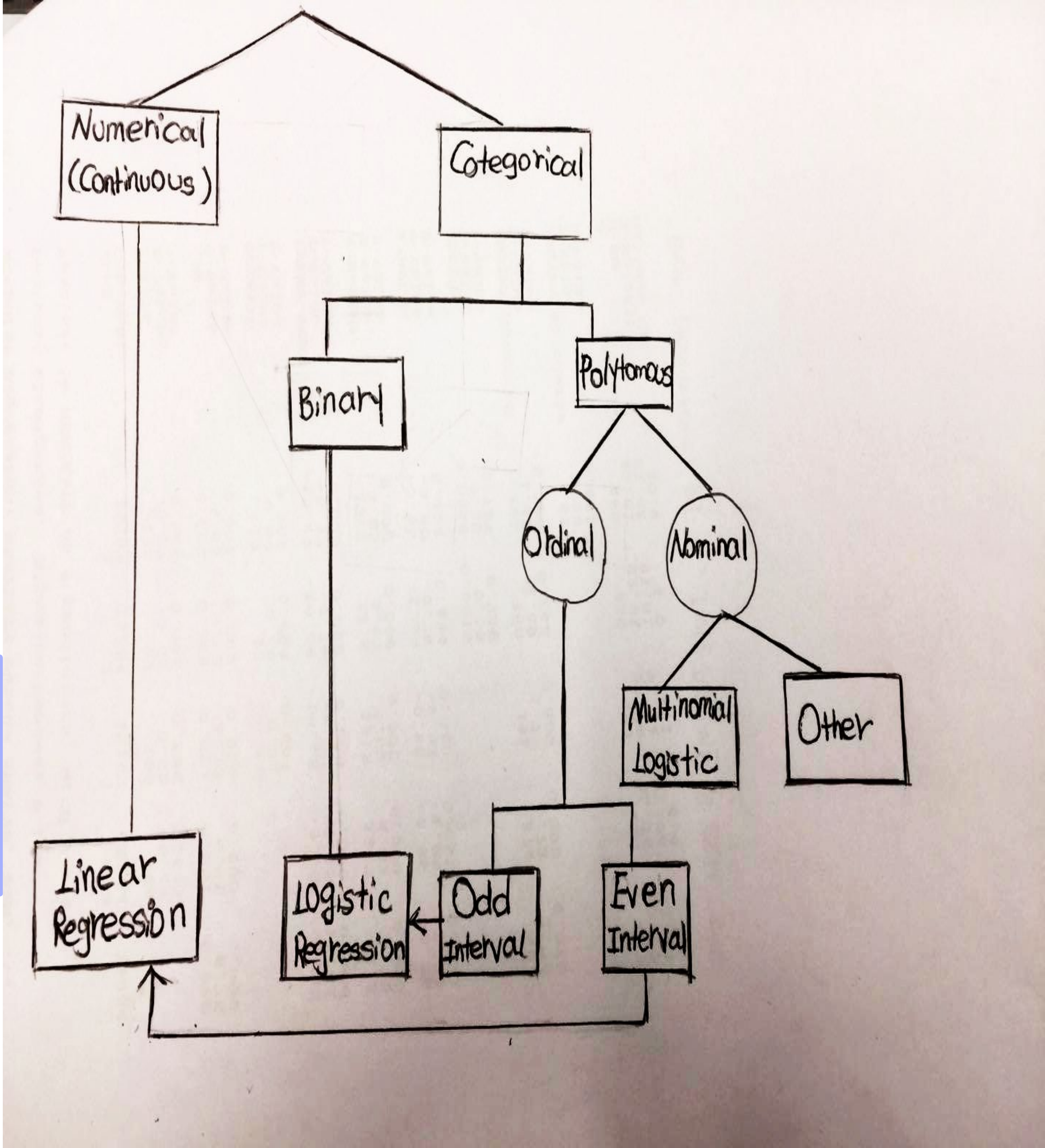
Ordinal with irregular intervals -> Logistic Regression

Ordinal with even intervals -> Logistic Regression OR (decision tree) OR Linear Regression (like Ryan's presentation on Predicting Map Difficulty with Dancing)

Nominal-> multinomial logit model

Linear or Logistic Regression Model?





WHY LOGISTIC REGRESSION IS NEEDED？

1. The residuals cannot be normally distributed (as the OLS model assumes), since they can only take on one of several values for each combination of level of the Independent Variables

2. The OLS model makes nonsensical predictions, since the DV is not continuous - e.g., it may predict that someone does something more than ‘all the time’. Something like 1.3 doesn’t make sense in a binary response.

3. For nominal DVs, the coding is completely arbitrary.

For Ordinal Data, a linear regression may make sense if they are evenly spaced.

How to predict a categorical response?

In statistics, logistic regression, or logit regression, or logit model[1] is a regression model where the dependent variable (DV) is categorical.

Assumptions for Logistic Regressions

Don’t Need

Doesn’t need to be linear ✗

Distribution doesn’t need to be normal ✗

Do Need:

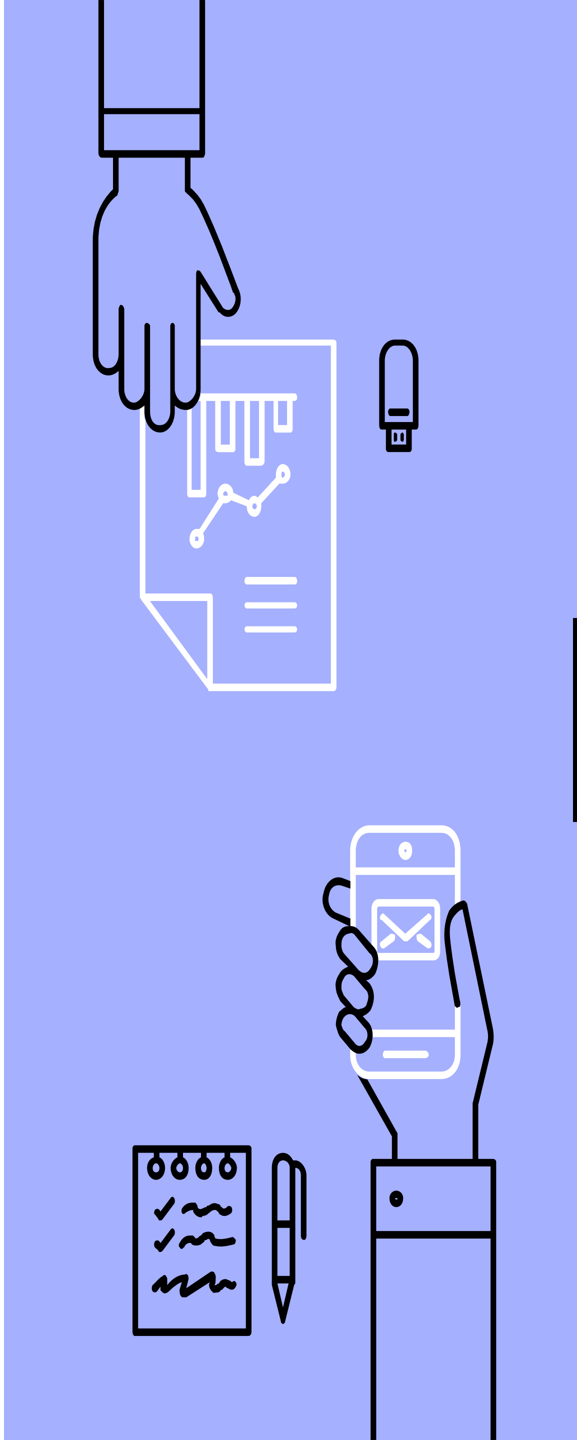
Binary or ordinal data ✓

No multicollinearity in variables ✓

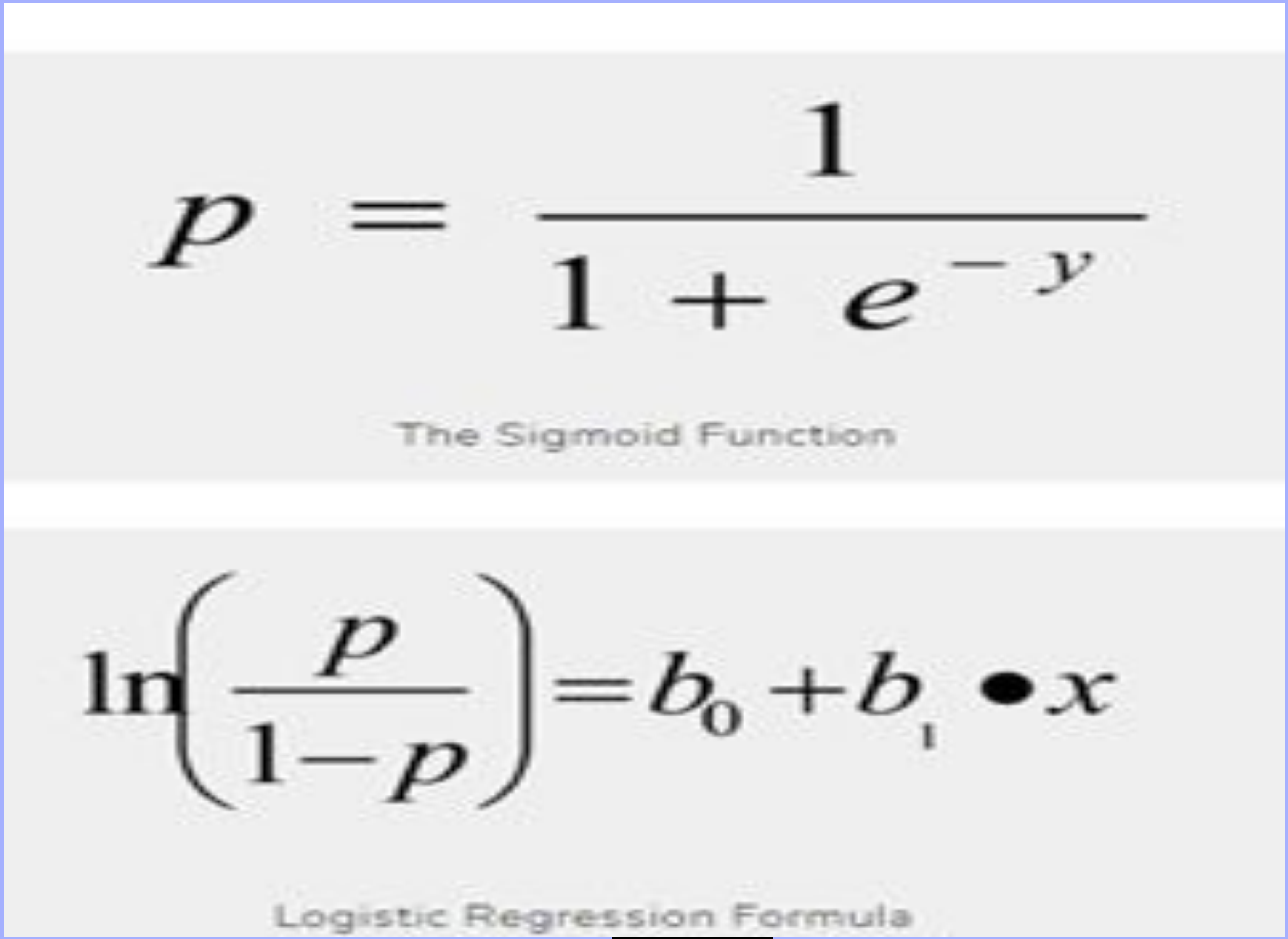
Independent observations (no matched data) ✓

Independent variables need to be linearly related to the log odds ✓

Large sample size ✓



Logistic Regression Formula:



$$p = \frac{1}{1 + e^{-y}}$$

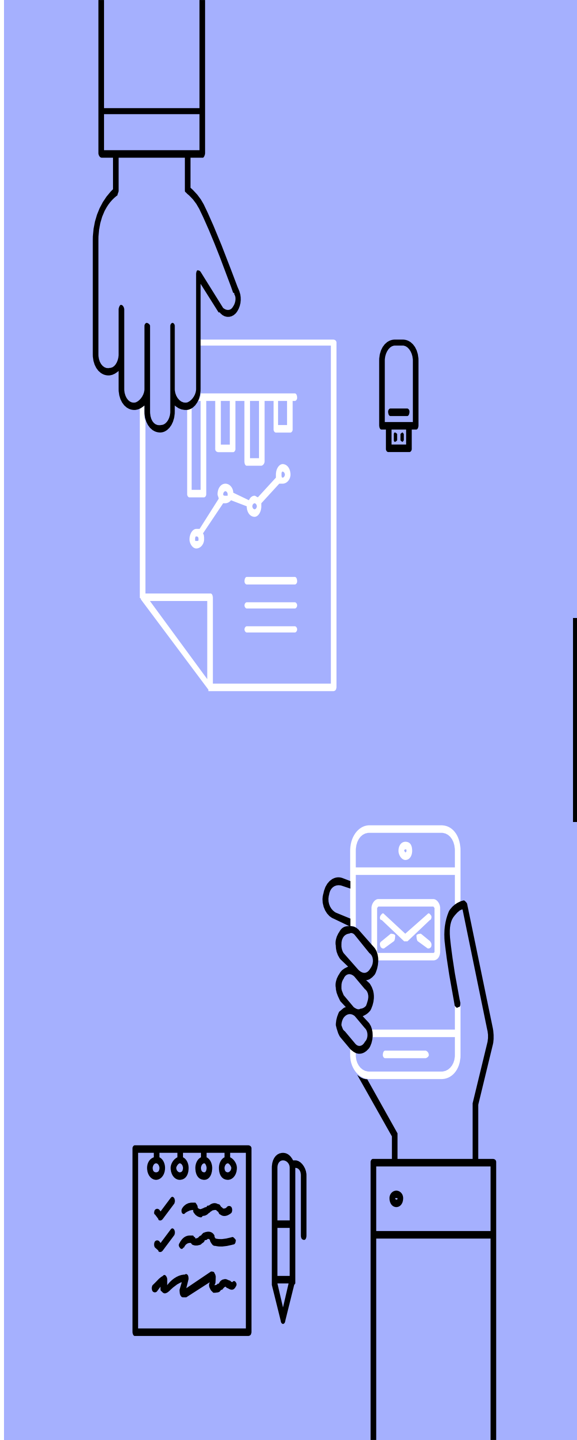
The Sigmoid Function

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \bullet x$$

Logistic Regression Formula

Binary

A binary logistic model is used to **estimate the probability of a binary response** based on one or more predictor (or independent) variables (features).

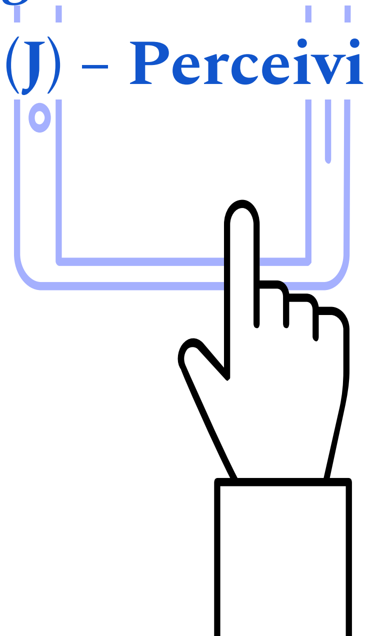


Case Study: Using forum posts to predict a person's personality

WHAT IS MBTI ?

The Myers Briggs Type Indicator (MBTI) is a personality classification system that divides individuals into 16 distinct personality types across 4 axis:

- Introversion (I) – Extroversion (E)
- Intuition (N) – Sensing (S)
- Thinking (T) – Feeling (F)
- Judging (J) – Perceiving (P)



16 Types

ISTJ
Inspector

ISFJ
Protector

INFJ
Counselor

INTJ
Mastermind

ISTP
Crafter

ISFP
Composer

INFP
Healer

INTP
Architect

ESTP
Promoter

ESFP
Performer

ENFP
Champion

ENTP
Inventor

ESTJ
Supervisor

ESFJ
Provider

ENFJ
Teacher

ENTJ
Fieldmarshal

youtopiaproject.com

The 20 Pros and Cons of Being an ENFP

by Vienna Kendall

of Dating an
ale

INFP Love: 4 Things To Know About INFPs to Prevent

Which Game of Thrones Character Are You?

by Aaron Y

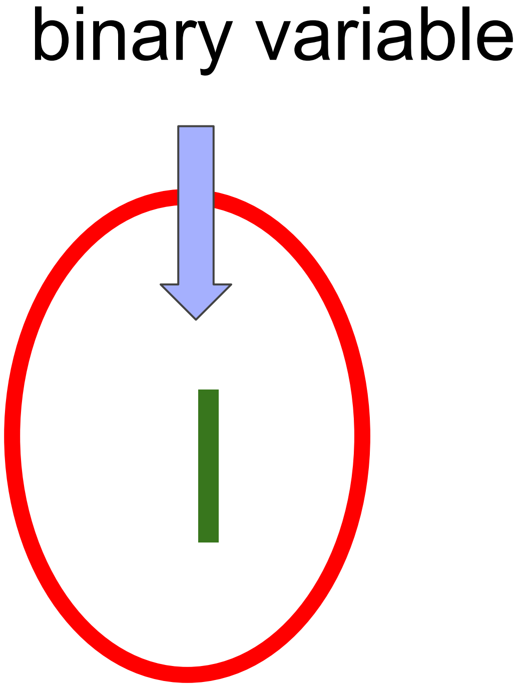
Game of Thrones Shipping MBTI Style: Romance Between

INFJ-INTJ: Ideal INF

First axis I VS. E

Introverted V.S **Extroverted**

First Indicator I VS. E


binary variable

I

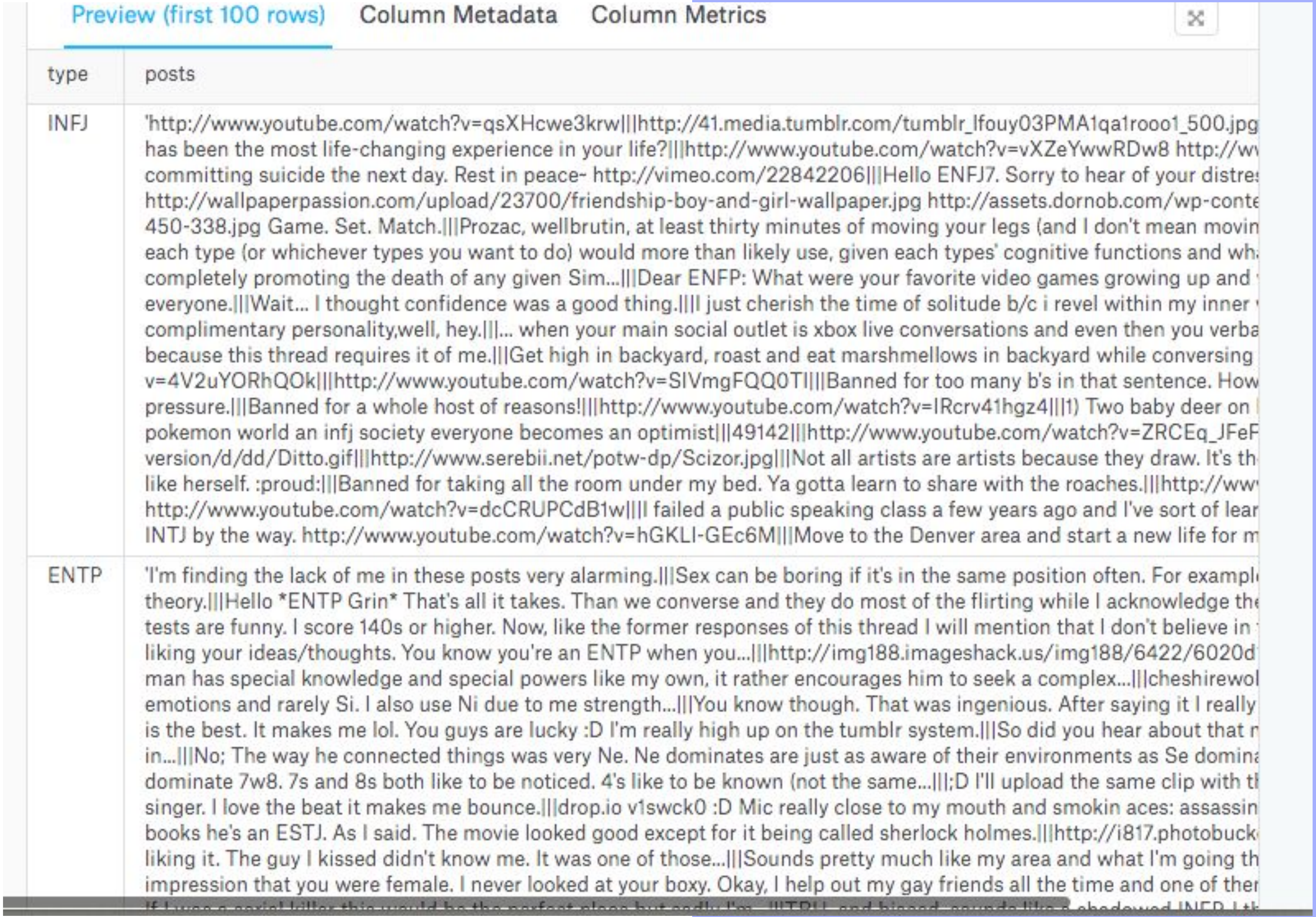
N F P

Do using certain words indicate Introverted/Extrovertedness?

Do specific words indicate a certain temperament/personality???

MBTI Dataset

https://www.kaggle.com/datasnaek/mbti-type/data



type	posts
INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg has been the most life-changing experience in your life?|||http://www.youtube.com/watch?v=vXZeYwwRDw8 http://ww committing suicide the next day. Rest in peace~ http://vimeo.com/22842206|||Hello ENFJ7. Sorry to hear of your distres http://wallpaperpassion.com/upload/23700/friendship-boy-and-girl-wallpaper.jpg http://assets.dornob.com/wp-conte 450-338.jpg Game. Set. Match.|||Prozac, wellbrutin, at least thirty minutes of moving your legs (and I don't mean movin each type (or whichever types you want to do) would more than likely use, given each types' cognitive functions and wh completely promoting the death of any given Sim...|||Dear ENFP: What were your favorite video games growing up and everyone.|||Wait... I thought confidence was a good thing.|||I just cherish the time of solitude b/c i revel within my inner complimentary personality,well, hey.|||... when your main social outlet is xbox live conversations and even then you verba because this thread requires it of me.|||Get high in backyard, roast and eat marshmellows in backyard while conversing v=4V2uYORhQOk|||http://www.youtube.com/watch?v=SlVmgFQQ0TI|||Banned for too many b's in that sentence. How pressure.|||Banned for a whole host of reasons!|||http://www.youtube.com/watch?v=IRcrv41hgz4|||1) Two baby deer on pokemon world an infj society everyone becomes an optimist|||49142|||http://www.youtube.com/watch?v=ZRCEq_JFeF version/d/dd/Ditto.gif|||http://www.serebii.net/potw-dp/Scizor.jpg|||Not all artists are artists because they draw. It's th like herself. :proud:|||Banned for taking all the room under my bed. Ya gotta learn to share with the roaches.|||http://ww http://www.youtube.com/watch?v=dcCRUPCdB1w|||I failed a public speaking class a few years ago and I've sort of lear INTJ by the way. http://www.youtube.com/watch?v=hGKLI-GEc6M|||Move to the Denver area and start a new life for m
ENTP	'I'm finding the lack of me in these posts very alarming.|||Sex can be boring if it's in the same position often. For exampl theory.|||Hello *ENTP Grin* That's all it takes. Than we converse and they do most of the flirting while I acknowledge the tests are funny. I score 140s or higher. Now, like the former responses of this thread I will mention that I don't believe in liking your ideas/thoughts. You know you're an ENTP when you...|||http://img188.imageshack.us/img188/6422/6020d man has special knowledge and special powers like my own, it rather encourages him to seek a complex...|||cheshirewol emotions and rarely Si. I also use Ni due to me strength...|||You know though. That was ingenious. After saying it I really is the best. It makes me lol. You guys are lucky :D I'm really high up on the tumblr system.|||So did you hear about that r in...|||No; The way he connected things was very Ne. Ne dominates are just as aware of their environments as Se domina dominate 7w8. 7s and 8s both like to be noticed. 4's like to be known (not the same...|||;D I'll upload the same clip with th singer. I love the beat it makes me bounce.|||drop.io v1swck0 :D Mic really close to my mouth and smokin aces: assassin books he's an ESTJ. As I said. The movie looked good except for it being called sherlock holmes.|||http://i817.photobuck liking it. The guy I kissed didn't know me. It was one of those...|||Sounds pretty much like my area and what I'm going th impression that you were female. I never looked at your boxy. Okay, I help out my gay friends all the time and one of ther If I was a serial killer this would be the perfect place but sadly I'm...|||TRU and biased, sounds like a shadowed INFP I th

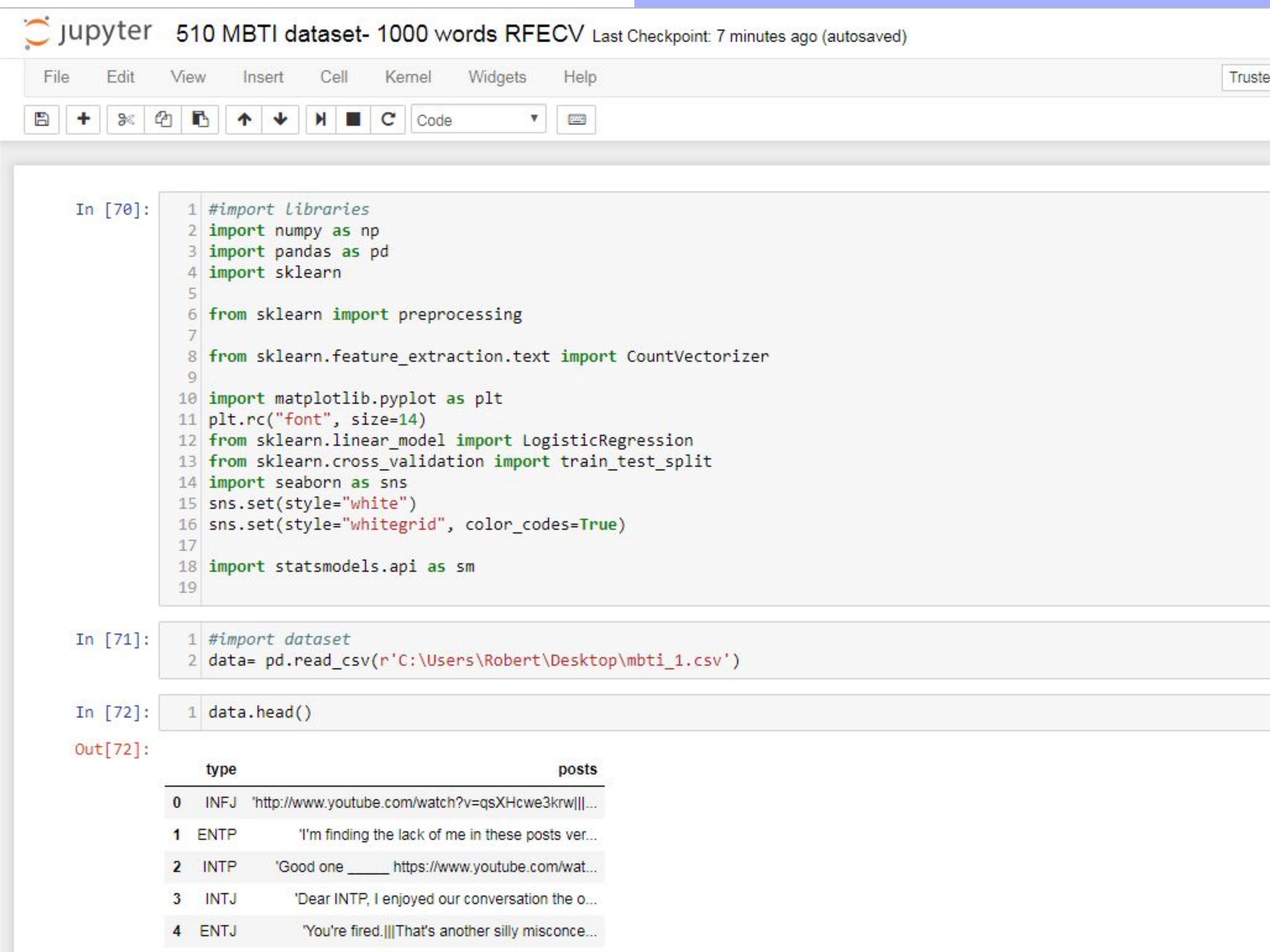
Do using certain words indicate a certain temperament?

Procedure

1. **Clean data**
   - Create a column for I vs. E (binary)
2. **Tokenize/vectorize dataset (bag of words)**
   - Count words & create giant matrix
3. **Run a logistic regression**
4. **Summarize data**

Case Study: MBTI

Python code



510 MBTI dataset- 1000 words RFECV Last Checkpoint: 7 minutes ago (autosaved)

In [70]:

```python
#import libraries
import numpy as np
import pandas as pd
import sklearn

from sklearn import preprocessing

from sklearn.feature_extraction.text import CountVectorizer

import matplotlib.pyplot as plt
plt.rc("font", size=14)
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split
import seaborn as sns
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)

import statsmodels.api as sm

```

In [71]:

```python
#import dataset
data= pd.read_csv(r'C:\Users\Robert\Desktop\mbti_1.csv')
```

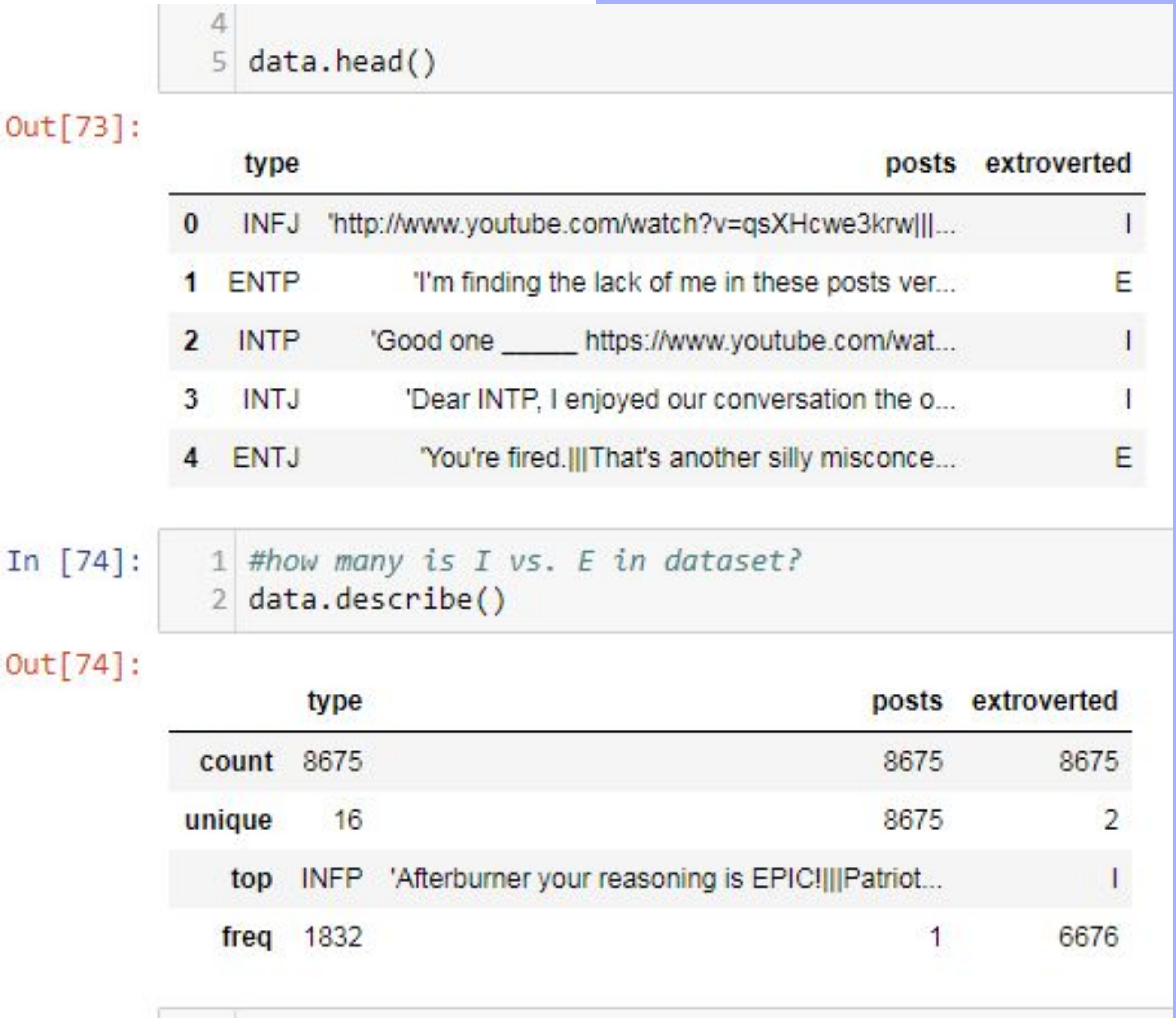
In [72]:

```python
data.head()
```

Out[72]:

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Our Data



```

data.head()
```

Out[73]:

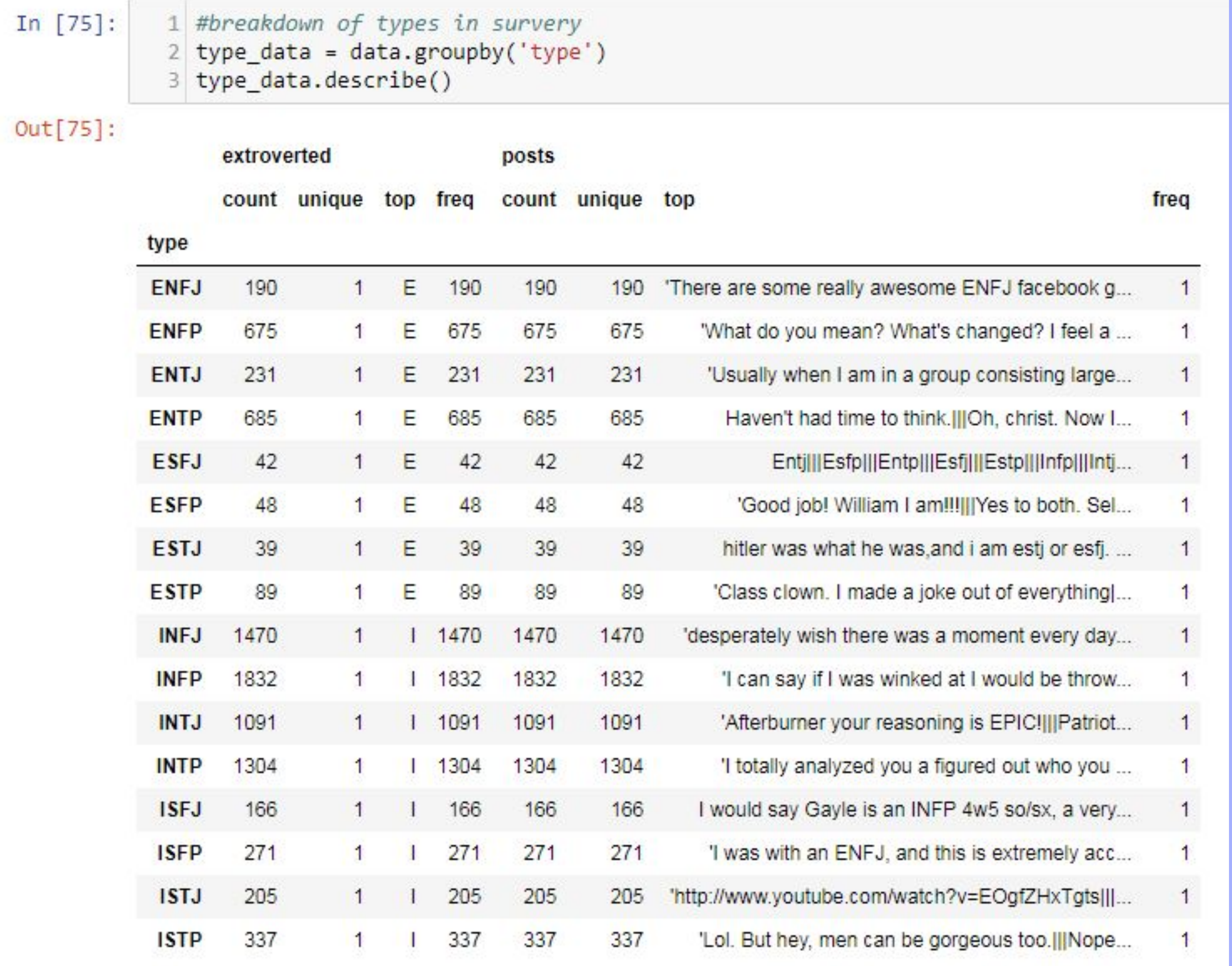
	type	posts	extroverted
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	I
1	ENTP	'I'm finding the lack of me in these posts ver...	E
2	INTP	'Good one _____ https://www.youtube.com/wat...	I
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	I
4	ENTJ	'You're fired.|||That's another silly misconce...	E

In [74]:

```
#how many is I vs. E in dataset?
data.describe()
```

Out[74]:

	type	posts	extroverted
count	8675	8675	8675
unique	16	8675	2
top	INFP	'Afterburner your reasoning is EPIC!|||Patriot...	I
freq	1832	1	6676

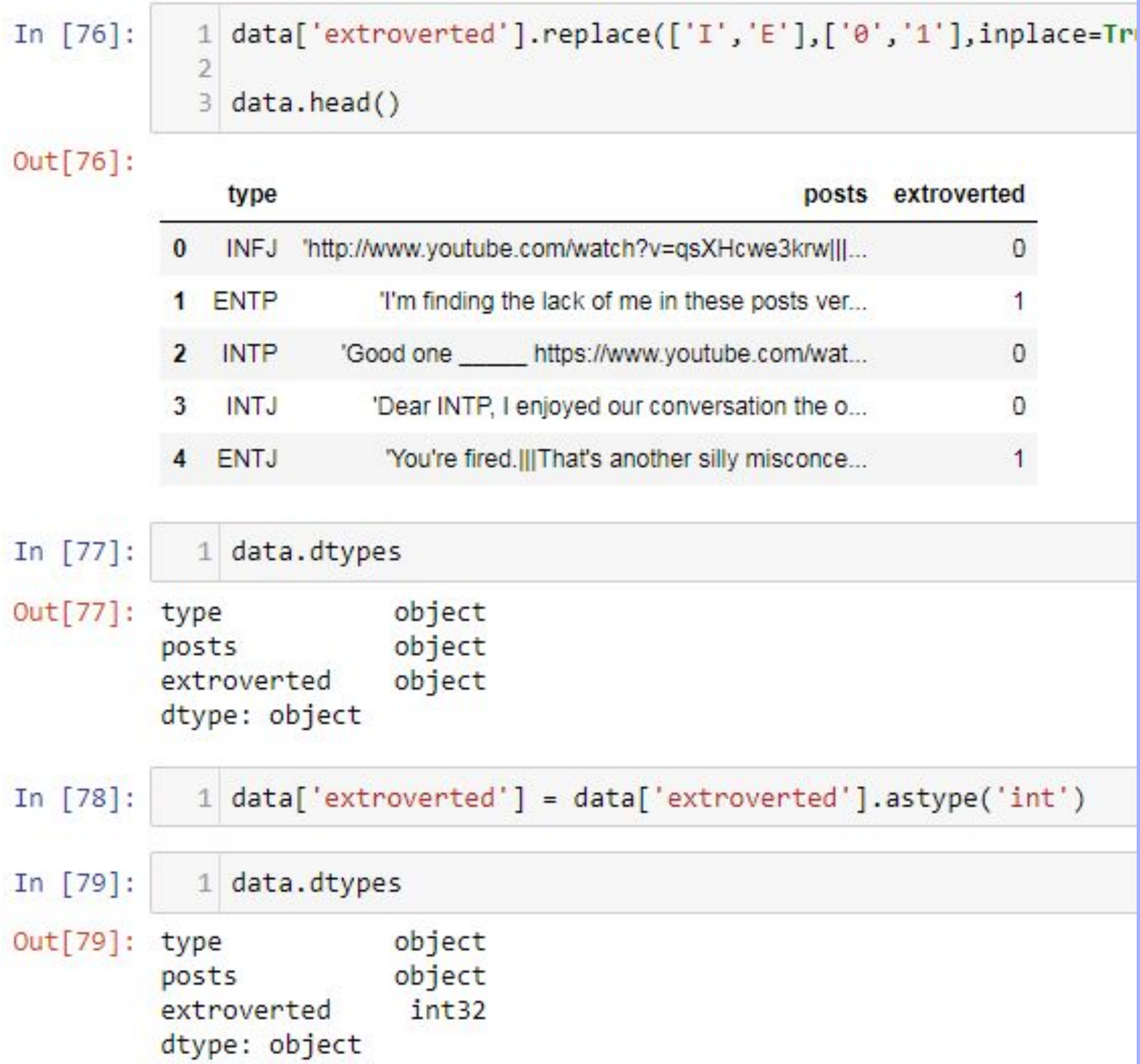


In [75]:
```
#breakdown of types in survery
type_data = data.groupby('type')
type_data.describe()
```

Out[75]:

	extroverted				posts			
	count	unique	top	freq	count	unique	top	freq
type								
ENFJ	190	1	E	190	190	190	'There are some really awesome ENFJ facebook g...	1
ENFP	675	1	E	675	675	675	'What do you mean? What's changed? I feel a ...	1
ENTJ	231	1	E	231	231	231	'Usually when I am in a group consisting large...	1
ENTP	685	1	E	685	685	685	Haven't had time to think.|||Oh, christ. Now I...	1
ESFJ	42	1	E	42	42	42	Entj|||Esfp|||Entp|||Esfj|||Estp|||Infp|||Intj...	1
ESFP	48	1	E	48	48	48	'Good job! William I am!!!|||Yes to both. Sel...	1
ESTJ	39	1	E	39	39	39	hitler was what he was,and i am estj or esfj. ...	1
ESTP	89	1	E	89	89	89	'Class clown. I made a joke out of everything|...	1
INFJ	1470	1	I	1470	1470	1470	'desperately wish there was a moment every day...	1
INFP	1832	1	I	1832	1832	1832	'I can say if I was winked at I would be throw...	1
INTJ	1091	1	I	1091	1091	1091	'Afterburner your reasoning is EPIC!|||Patriot...	1
INTP	1304	1	I	1304	1304	1304	'I totally analyzed you a figured out who you ...	1
ISFJ	166	1	I	166	166	166	I would say Gayle is an INFP 4w5 so/sx, a very...	1
ISFP	271	1	I	271	271	271	'I was with an ENFJ, and this is extremely acc...	1
ISTJ	205	1	I	205	205	205	'http://www.youtube.com/watch?v=EOgfZHxTgts|||...	1
ISTP	337	1	I	337	337	337	'Lol. But hey, men can be gorgeous too.|||Nope...	1

Make Binary Column



In [76]:
```
data['extroverted'].replace(['I','E'],['0','1'],inplace=Tr

data.head()
```

Out[76]:

	type	posts	extroverted
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	0
1	ENTP	'I'm finding the lack of me in these posts ver...	1
2	INTP	'Good one _____ https://www.youtube.com/wat...	0
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	0
4	ENTJ	'You're fired.|||That's another silly misconce...	1

In [77]:
```
data.dtypes
```

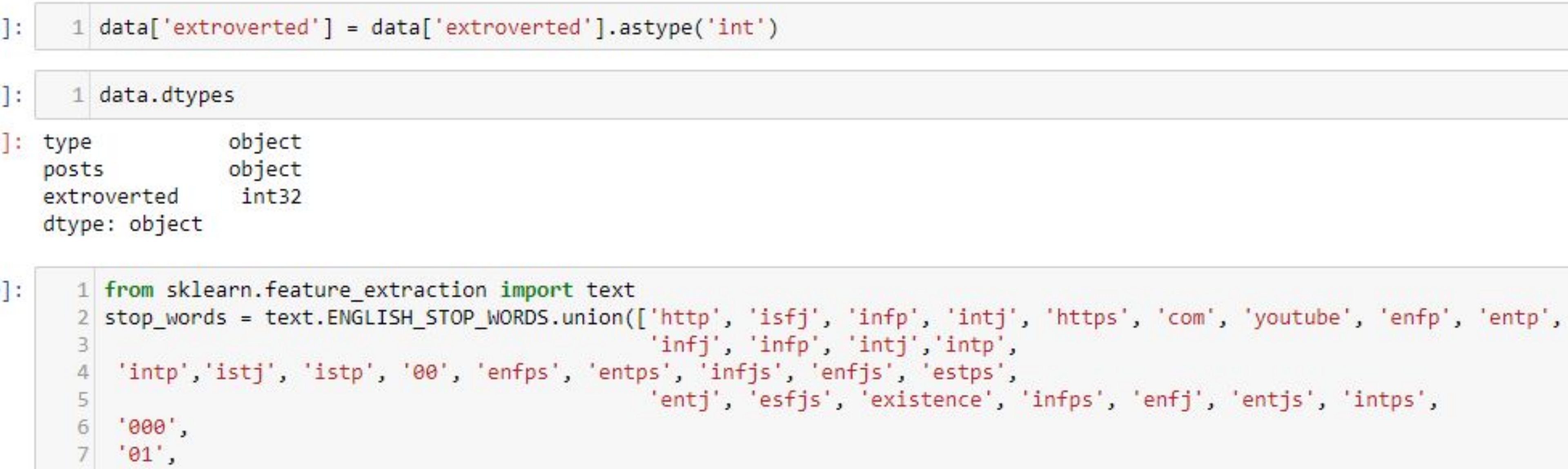
Out[77]:
```
type            object
posts           object
extroverted     object
dtype: object
```

In [78]:
```
data['extroverted'] = data['extroverted'].astype('int')
```

In [79]:
```
data.dtypes
```

Out[79]:
```
type            object
posts           object
extroverted      int32
dtype: object
```

Create Stopwords



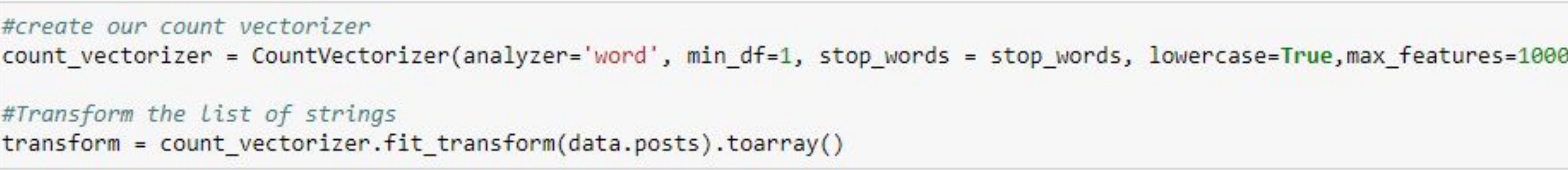
```
data['extroverted'] = data['extroverted'].astype('int')
```

```
data.dtypes
```

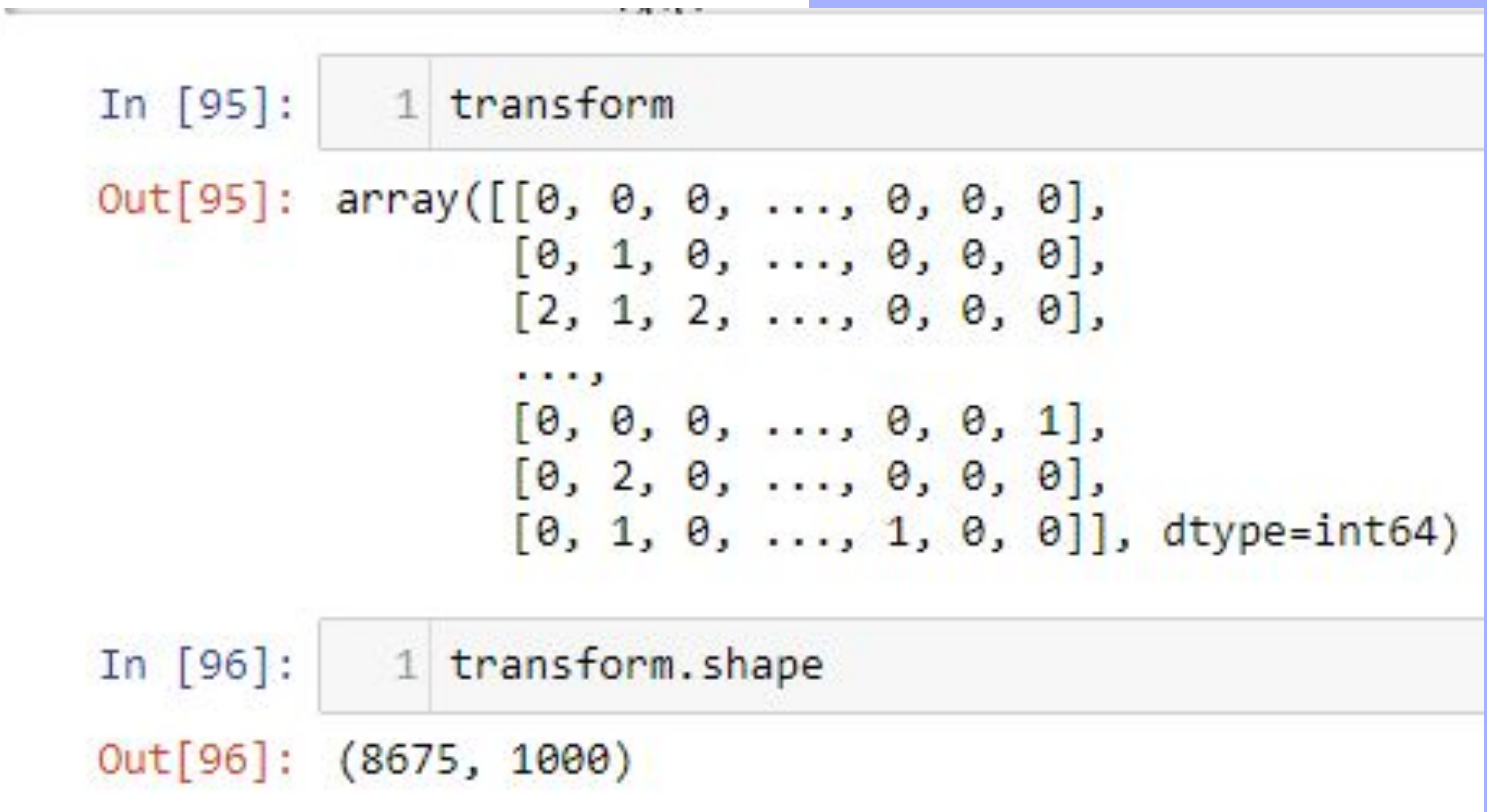
```
type            object
posts           object
extroverted      int32
dtype: object
```

```
from sklearn.feature_extraction import text
stop_words = text.ENGLISH_STOP_WORDS.union(['http', 'isfj', 'infp', 'intj', 'https', 'com', 'youtube', 'enfp', 'entp',
                                            'infj', 'infp', 'intj','intp',
 'intp','istj', 'istp', '00', 'enfps', 'entps', 'infjs', 'enfjs', 'estps',
                                           'entj', 'esfjs', 'existence', 'infps', 'enfj', 'entjs', 'intps',
 '000',
 '01',
```

Count Vectorizer

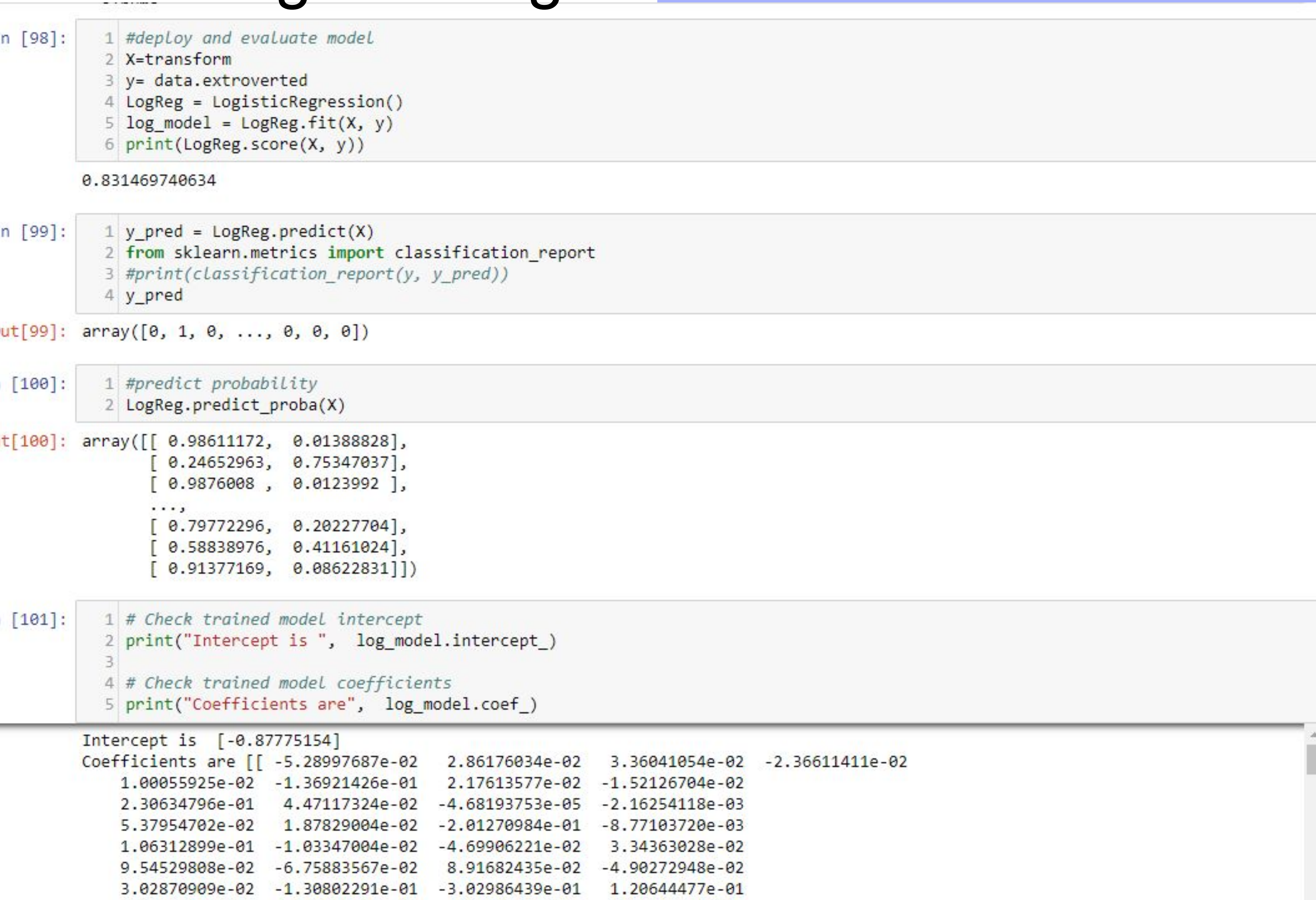


```
#create our count vectorizer
count_vectorizer = CountVectorizer(analyzer='word', min_df=1, stop_words = stop_words, lowercase=True,max_features=1000

#Transform the list of strings
transform = count_vectorizer.fit_transform(data.posts).toarray()
```



```
In [95]:  transform

Out[95]: array([[0, 0, 0, ..., 0, 0, 0],
                [0, 1, 0, ..., 0, 0, 0],
                [2, 1, 2, ..., 0, 0, 0],
                ...,
                [0, 0, 0, ..., 0, 0, 1],
                [0, 2, 0, ..., 0, 0, 0],
                [0, 1, 0, ..., 1, 0, 0]], dtype=int64)

In [96]:  transform.shape

Out[96]: (8675, 1000)
```

Initial Logistic Regression



n [98]:

```
#deploy and evaluate model
X=transform
y= data.extroverted
LogReg = LogisticRegression()
log_model = LogReg.fit(X, y)
print(LogReg.score(X, y))
```

```
0.831469740634
```

n [99]:

```
y_pred = LogReg.predict(X)
from sklearn.metrics import classification_report
#print(classification_report(y, y_pred))
y_pred
```

ut[99]:

```
array([0, 1, 0, ..., 0, 0, 0])
```

[100]:

```
#predict probability
LogReg.predict_proba(X)
```

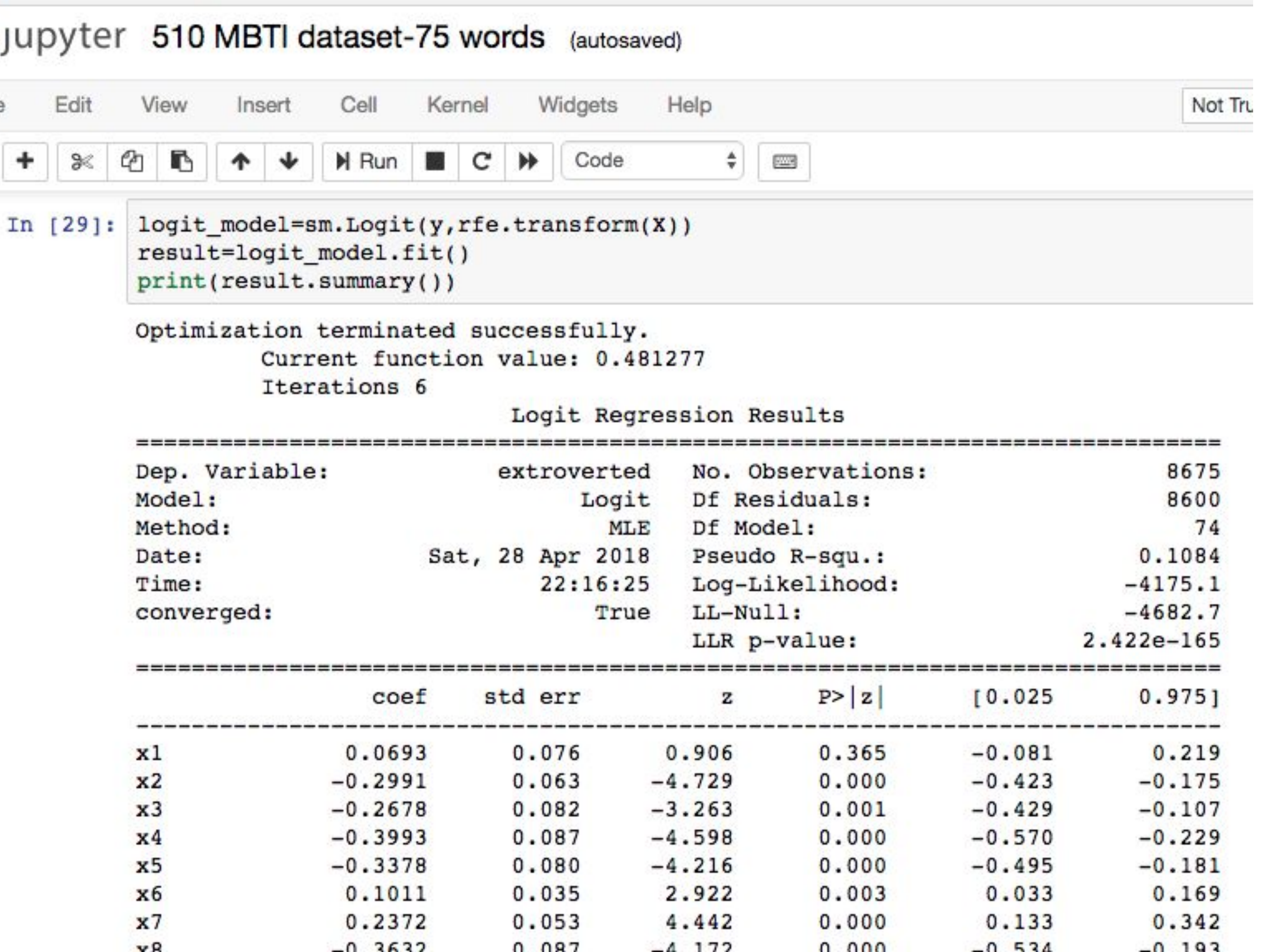
t[100]:

```
array([[ 0.98611172,  0.01388828],
       [ 0.24652963,  0.75347037],
       [ 0.9876008 ,  0.0123992 ],
       ...,
       [ 0.79772296,  0.20227704],
       [ 0.58838976,  0.41161024],
       [ 0.91377169,  0.08622831]])
```

[101]:

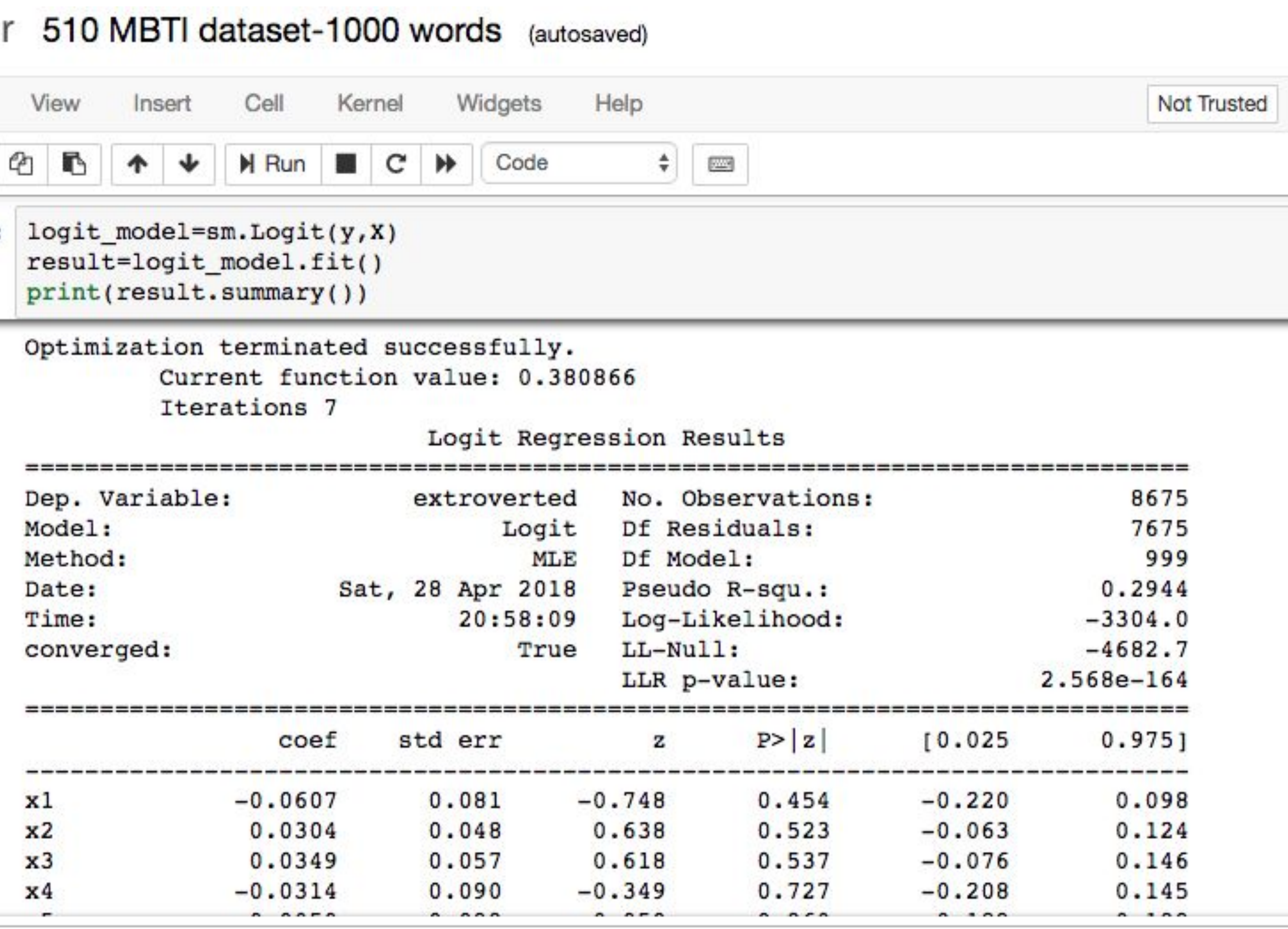
```
# Check trained model intercept
print("Intercept is ",  log_model.intercept_)

# Check trained model coefficients
print("Coefficients are",  log_model.coef_)
```

```
Intercept is  [-0.87775154]
Coefficients are [[ -5.28997687e-02   2.86176034e-02   3.36041054e-02  -2.36611411e-02
    1.00055925e-02  -1.36921426e-01   2.17613577e-02  -1.52126704e-02
    2.30634796e-01   4.47117324e-02  -4.68193753e-05  -2.16254118e-03
    5.37954702e-02   1.87829004e-02  -2.01270984e-01  -8.77103720e-03
    1.06312899e-01  -1.03347004e-02  -4.69906221e-02   3.34363028e-02
    9.54529808e-02  -6.75883567e-02   8.91682435e-02  -4.90272948e-02
    3.02870909e-02  -1.30802291e-01  -3.02986439e-01   1.20644477e-01
```


jupyter 510 MBTI dataset-75 words (autosaved)
e Edit View Insert Cell Kernel Widgets Help
Not Tru
Run
Code

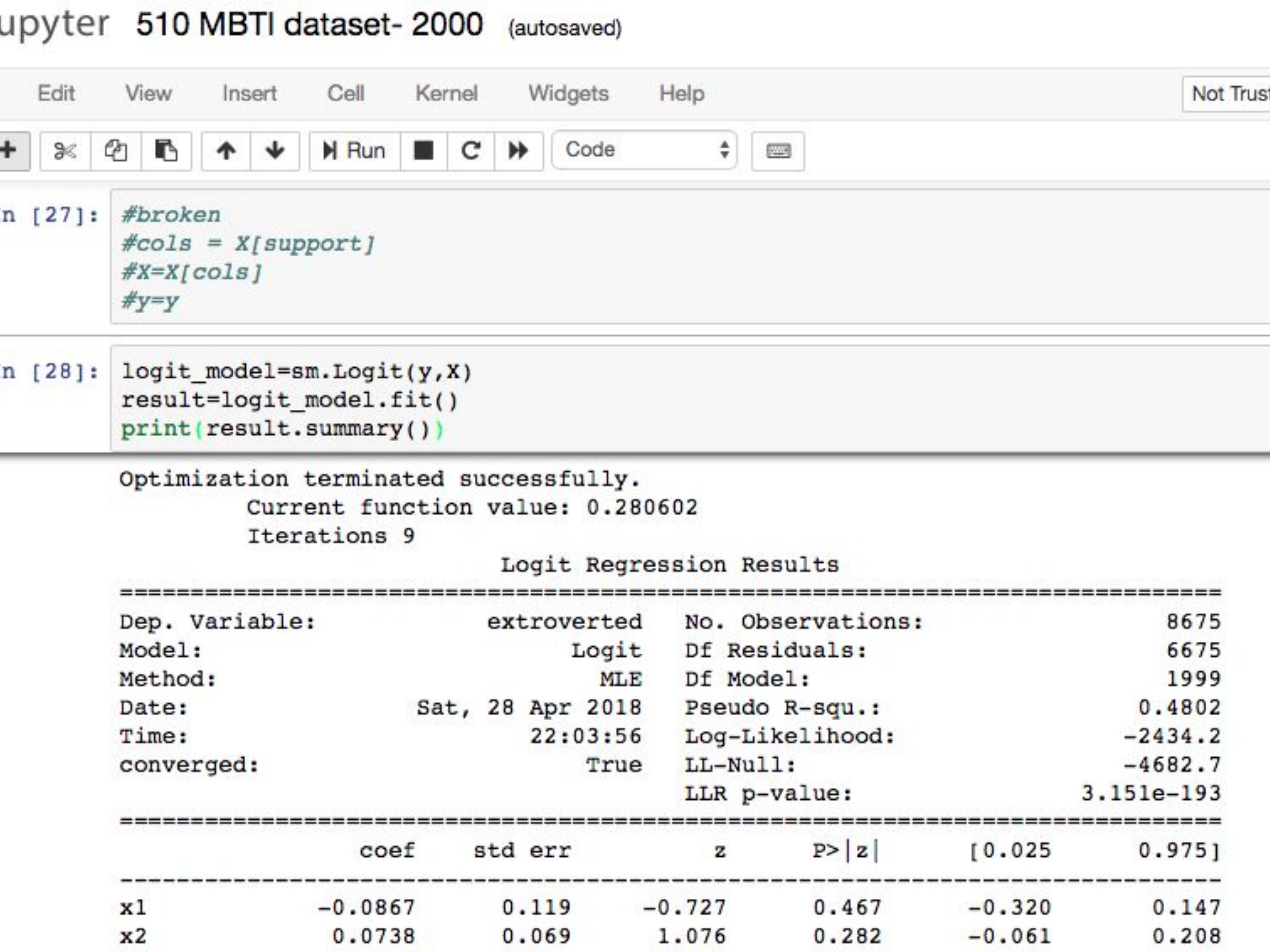
```
In [29]: logit_model=sm.Logit(y,rfe.transform(X))
         result=logit_model.fit()
         print(result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.481277
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:            extroverted   No. Observations:                 8675
Model:                          Logit   Df Residuals:                     8600
Method:                           MLE   Df Model:                           74
Date:                Sat, 28 Apr 2018   Pseudo R-squ.:                  0.1084
Time:                        22:16:25   Log-Likelihood:                -4175.1
converged:                       True   LL-Null:                       -4682.7
                                        LLR p-value:                 2.422e-165
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0693      0.076      0.906      0.365      -0.081       0.219
x2            -0.2991      0.063     -4.729      0.000      -0.423      -0.175
x3            -0.2678      0.082     -3.263      0.001      -0.429      -0.107
x4            -0.3993      0.087     -4.598      0.000      -0.570      -0.229
x5            -0.3378      0.080     -4.216      0.000      -0.495      -0.181
x6             0.1011      0.035      2.922      0.003       0.033       0.169
x7             0.2372      0.053      4.442      0.000       0.133       0.342
x8            -0.3632      0.087     -4.172      0.000      -0.534      -0.193
```



```
logit_model=sm.Logit(y,X)
result=logit_model.fit()
print(result.summary())
```

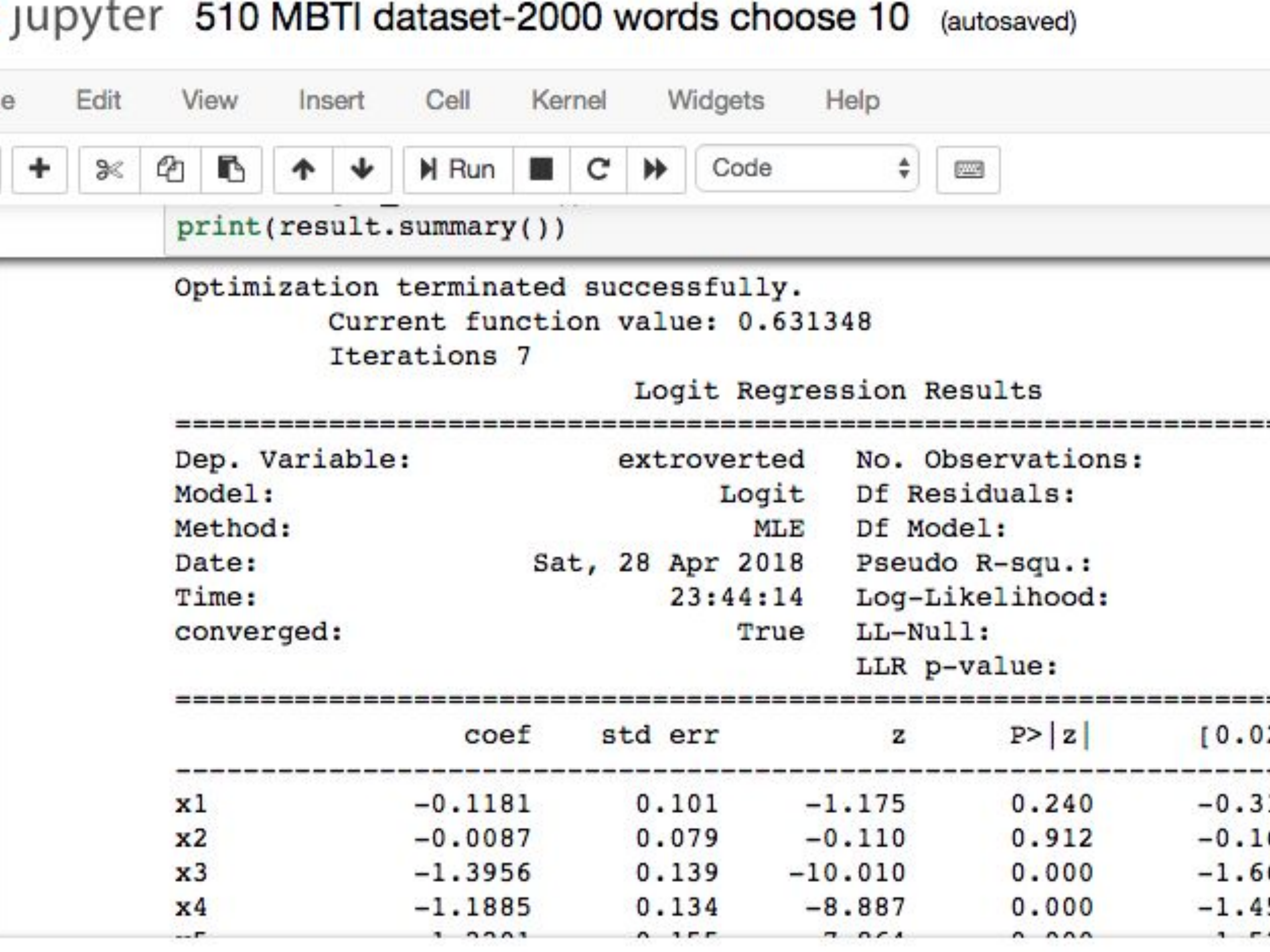
```
Optimization terminated successfully.
         Current function value: 0.380866
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:            extroverted   No. Observations:                 8675
Model:                          Logit   Df Residuals:                     7675
Method:                           MLE   Df Model:                          999
Date:                Sat, 28 Apr 2018   Pseudo R-squ.:                  0.2944
Time:                        20:58:09   Log-Likelihood:                -3304.0
converged:                       True   LL-Null:                       -4682.7
                                        LLR p-value:                2.568e-164
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0607      0.081     -0.748      0.454      -0.220       0.098
x2             0.0304      0.048      0.638      0.523      -0.063       0.124
x3             0.0349      0.057      0.618      0.537      -0.076       0.146
x4            -0.0314      0.090     -0.349      0.727      -0.208       0.145
```



```
#broken
#cols = X[support]
#X=X[cols]
#y=y
```

```
logit_model=sm.Logit(y,X)
result=logit_model.fit()
print(result.summary())
```

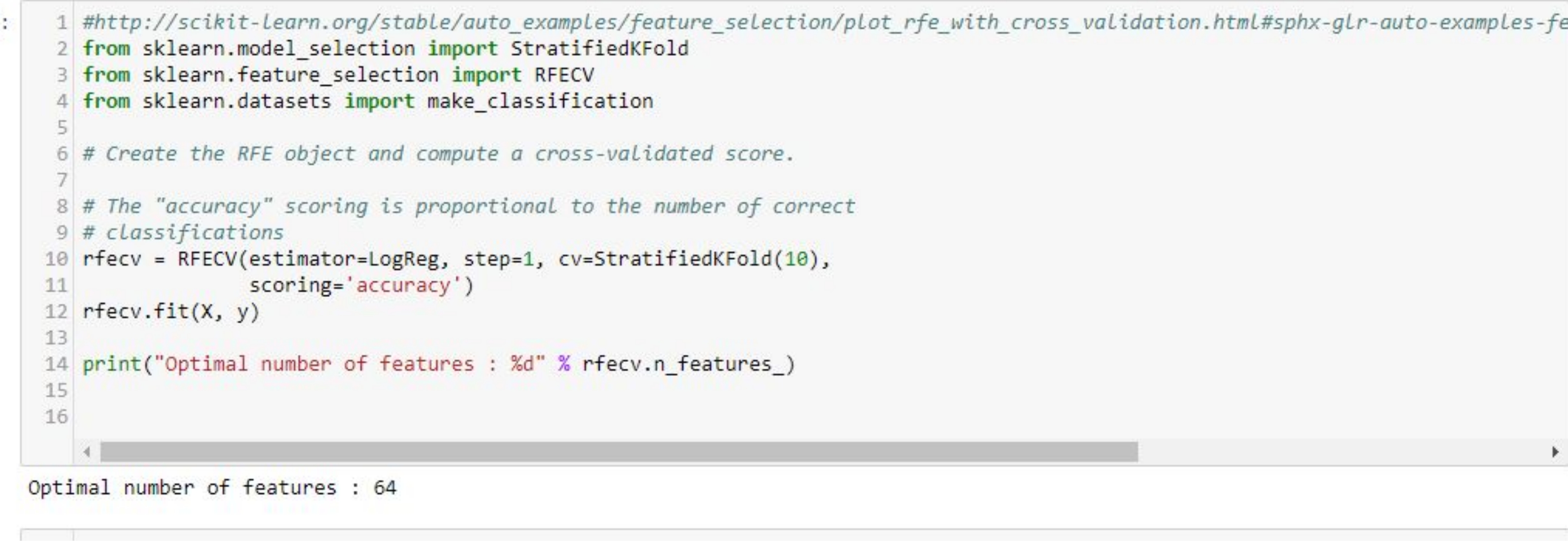
```
Optimization terminated successfully.
         Current function value: 0.280602
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:            extroverted   No. Observations:                 8675
Model:                          Logit   Df Residuals:                     6675
Method:                           MLE   Df Model:                         1999
Date:            Sat, 28 Apr 2018   Pseudo R-squ.:                  0.4802
Time:                    22:03:56   Log-Likelihood:                -2434.2
converged:                       True   LL-Null:                       -4682.7
                                        LLR p-value:                 3.151e-193
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0867      0.119     -0.727      0.467      -0.320       0.147
x2             0.0738      0.069      1.076      0.282      -0.061       0.208
```



```
print(result.summary())
```

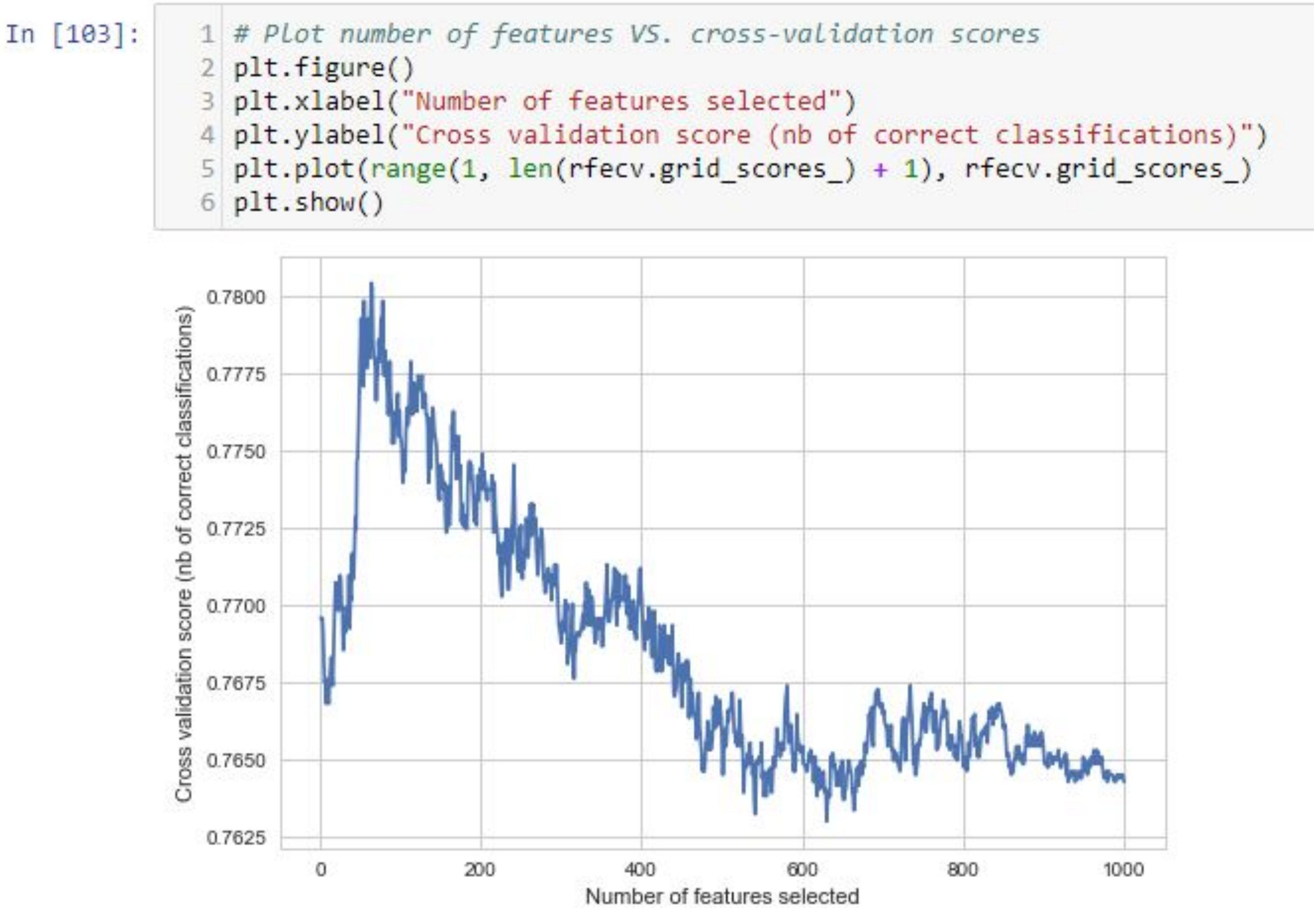
```
Optimization terminated successfully.
         Current function value: 0.631348
         Iterations 7
                          Logit Regression Results
==============================================================
Dep. Variable:            extroverted   No. Observations:
Model:                          Logit   Df Residuals:
Method:                           MLE   Df Model:
Date:              Sat, 28 Apr 2018   Pseudo R-squ.:
Time:                       23:44:14   Log-Likelihood:
converged:                      True   LL-Null:
                                       LLR p-value:
==============================================================
                 coef    std err          z      P>|z|      [0.0
--------------------------------------------------------------
x1            -0.1181      0.101     -1.175      0.240      -0.3
x2            -0.0087      0.079     -0.110      0.912      -0.1
x3            -1.3956      0.139    -10.010      0.000      -1.6
x4            -1.1885      0.134     -8.887      0.000      -1.4
```

Using Recursive Feature Selection with Cross Validation



```
#http://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html#sphx-glr-auto-examples-fe
from sklearn.model_selection import StratifiedKFold
from sklearn.feature_selection import RFECV
from sklearn.datasets import make_classification

# Create the RFE object and compute a cross-validated score.

# The "accuracy" scoring is proportional to the number of correct
# classifications
rfecv = RFECV(estimator=LogReg, step=1, cv=StratifiedKFold(10),
              scoring='accuracy')
rfecv.fit(X, y)

print("Optimal number of features : %d" % rfecv.n_features_)
```

```
Optimal number of features : 64
```

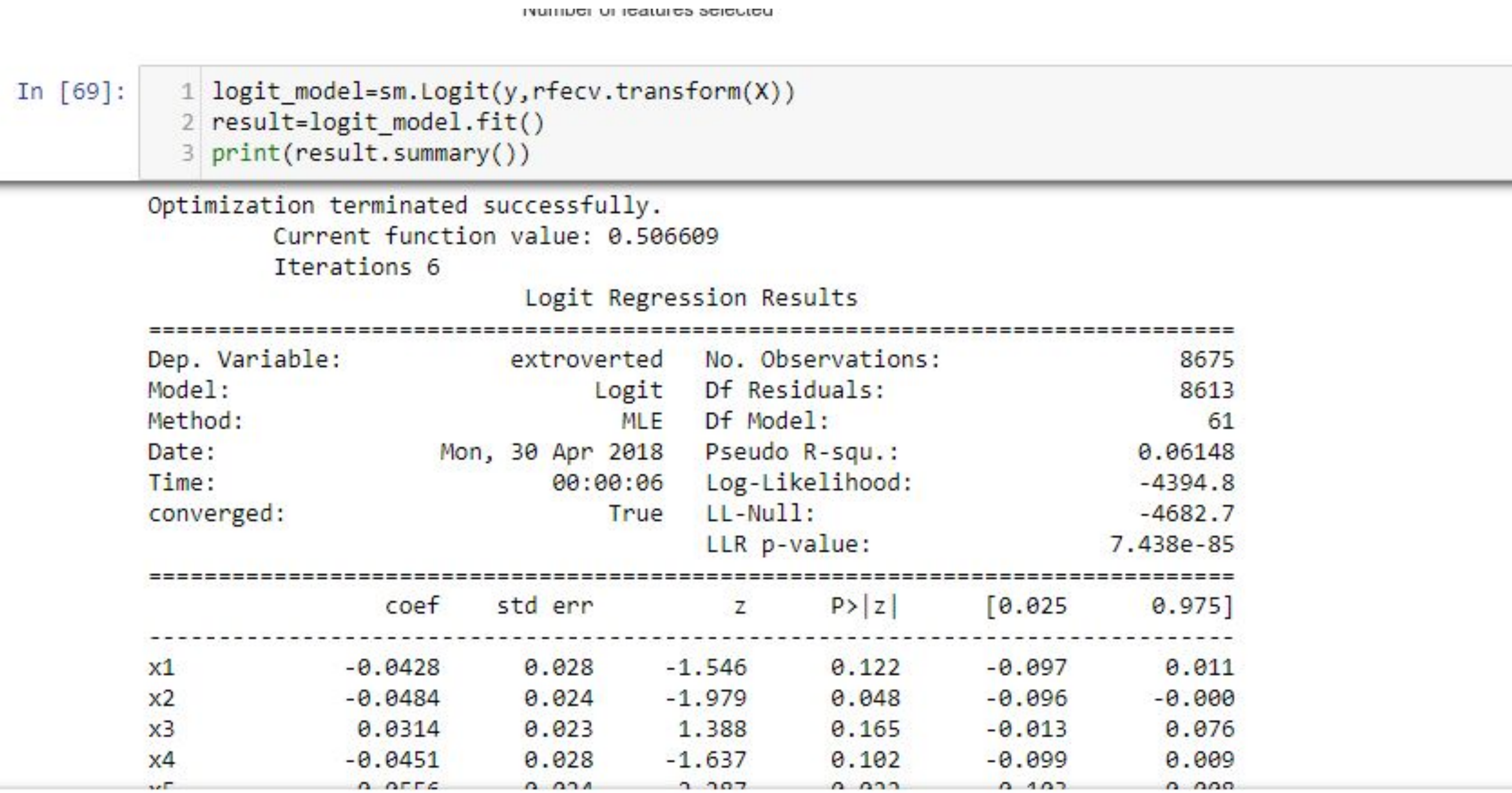


In [103]:

```python
# Plot number of features VS. cross-validation scores
plt.figure()
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score (nb of correct classifications)")
plt.plot(range(1, len(rfecv.grid_scores_) + 1), rfecv.grid_scores_)
plt.show()
```

Cross validation score (nb of correct classifications)
0.7800
0.7775
0.7750
0.7725
0.7700
0.7675
0.7650
0.7625
0
200
400
600
800
1000
Number of features selected

Using RFECV (Recursive Feature Elimination Cross Validation) 1000 Words

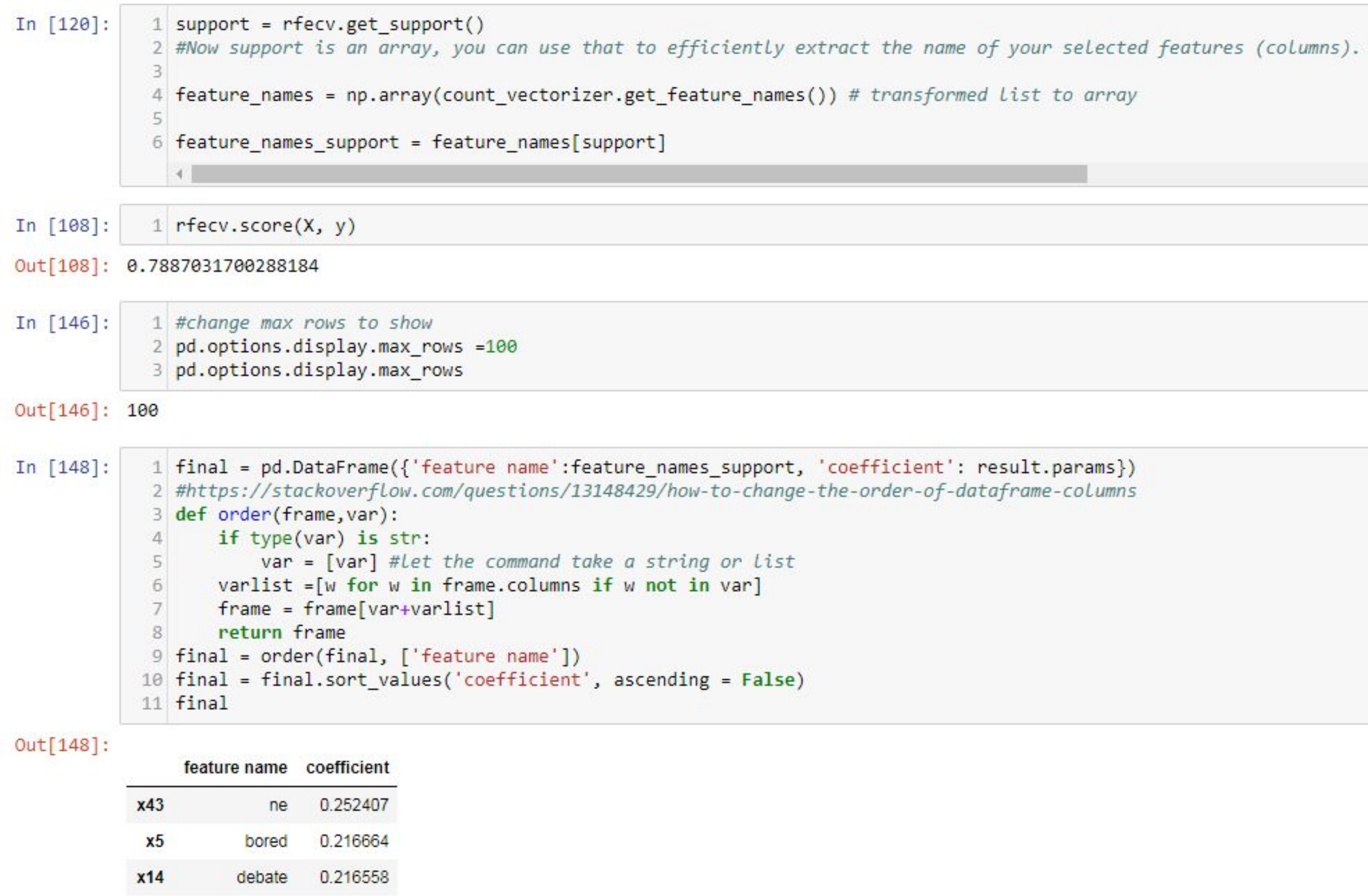


Number of features selected

In [69]:
```
logit_model=sm.Logit(y,rfecv.transform(X))
result=logit_model.fit()
print(result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.506609
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:            extroverted   No. Observations:                 8675
Model:                          Logit   Df Residuals:                     8613
Method:                           MLE   Df Model:                           61
Date:                Mon, 30 Apr 2018   Pseudo R-squ.:                 0.06148
Time:                        00:00:06   Log-Likelihood:                -4394.8
converged:                       True   LL-Null:                       -4682.7
                                        LLR p-value:                 7.438e-85
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0428      0.028     -1.546      0.122      -0.097       0.011
x2            -0.0484      0.024     -1.979      0.048      -0.096      -0.000
x3             0.0314      0.023      1.388      0.165      -0.013       0.076
x4            -0.0451      0.028     -1.637      0.102      -0.099       0.009
```

RFECV scores and feature names



```
In [120]:
support = rfecv.get_support()
#Now support is an array, you can use that to efficiently extract the name of your selected features (columns).

feature_names = np.array(count_vectorizer.get_feature_names()) # transformed list to array

feature_names_support = feature_names[support]
```

```
In [108]:
rfecv.score(X, y)
```

```
Out[108]: 0.7887031700288184
```

```
In [146]:
#change max rows to show
pd.options.display.max_rows =100
pd.options.display.max_rows
```

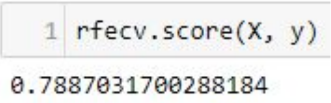
```
Out[146]: 100
```

```
In [148]:
final = pd.DataFrame({'feature name':feature_names_support, 'coefficient': result.params})
#https://stackoverflow.com/questions/13148429/how-to-change-the-order-of-dataframe-columns
def order(frame,var):
    if type(var) is str:
        var = [var] #let the command take a string or list
    varlist =[w for w in frame.columns if w not in var]
    frame = frame[var+varlist]
    return frame
final = order(final, ['feature name'])
final = final.sort_values('coefficient', ascending = False)
final
```

Out[148]:

	feature name	coefficient
x43	ne	0.252407
x5	bored	0.216664
x14	debate	0.216558

Conclusion:



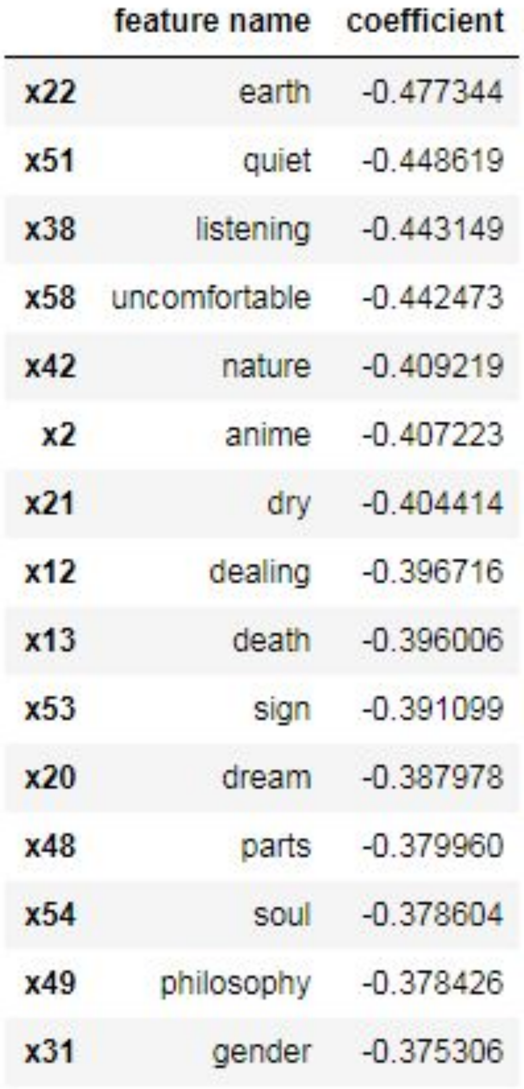
```
rfecv.score(X, y)
```

```
0.7887031700288184
```

Words used in posts can be used to predict temperament (at least introverted/extrovertedness)

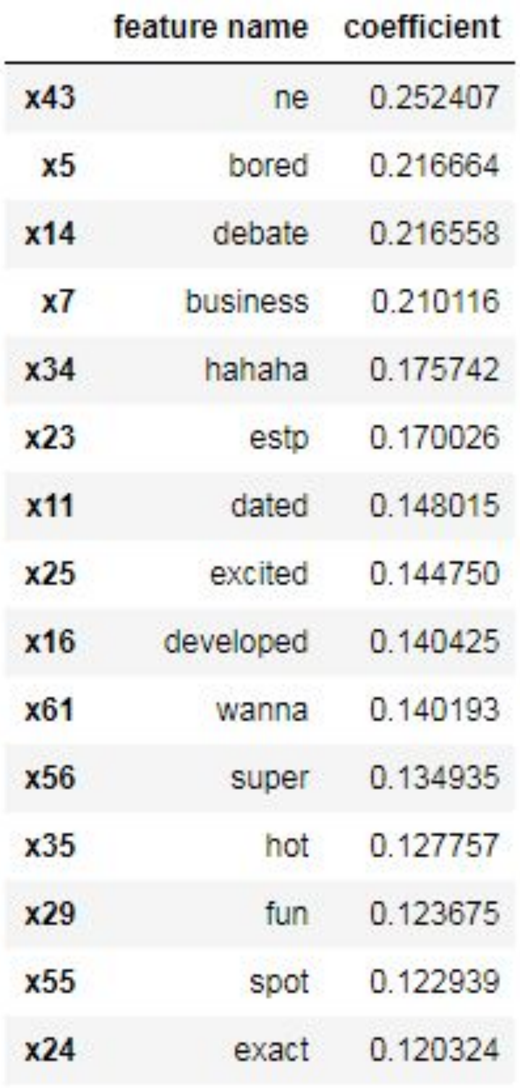
Top 15

Introverted Words



	feature name	coefficient
x22	earth	-0.477344
x51	quiet	-0.448619
x38	listening	-0.443149
x58	uncomfortable	-0.442473
x42	nature	-0.409219
x2	anime	-0.407223
x21	dry	-0.404414
x12	dealing	-0.396716
x13	death	-0.396006
x53	sign	-0.391099
x20	dream	-0.387978
x48	parts	-0.379960
x54	soul	-0.378604
x49	philosophy	-0.378426
x31	gender	-0.375306

Extroverted Words



	feature name	coefficient
x43	ne	0.252407
x5	bored	0.216664
x14	debate	0.216558
x7	business	0.210116
x34	hahaha	0.175742
x23	estp	0.170026
x11	dated	0.148015
x25	excited	0.144750
x16	developed	0.140425
x61	wanna	0.140193
x56	super	0.134935
x35	hot	0.127757
x29	fun	0.123675
x55	spot	0.122939
x24	exact	0.120324

Next Steps

- More Data
- Reduction of variables using lemmatization
- Try using ngrams
- Compare performance of other models/classifiers
- Look at each other axis N/S, F/T, J/P

Multinomial Models

Create multiple dummy regressions, each calculated simultaneously.

For multinomial models (more than 1 possible response) ????

https://www.mathworks.com/help/stats/multinomial-models-for-nominal-responses.html

http://amunategui.github.io/multinomial-neuralnetworks-walkthrough/

https://www.theanalysisfactor.com/logistic-regression-models-for-multinomial-and-ordinal-variables/

R walkthrough

References

https://en.wikipedia.org/wiki/Ordinal_regression

Ordinal Responses

Maybe an example? Which factor effects quality most? (Highest correlation) (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)

https://en.wikipedia.org/wiki/Ordered_logit

Example of ordered regression problems:

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

In SPSS

https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/categories/idh_catr.htm

General information on categorical data

http://www.ucd.ie/statdept/classpages/categorical_data_analysis/cda1.pdf

References for R, SAS, & Python

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

R tutorial on running a logistic regression:

https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

R tutorial on building a logistic regression from scratch:

https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/

Python walkthrough:
https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a

https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

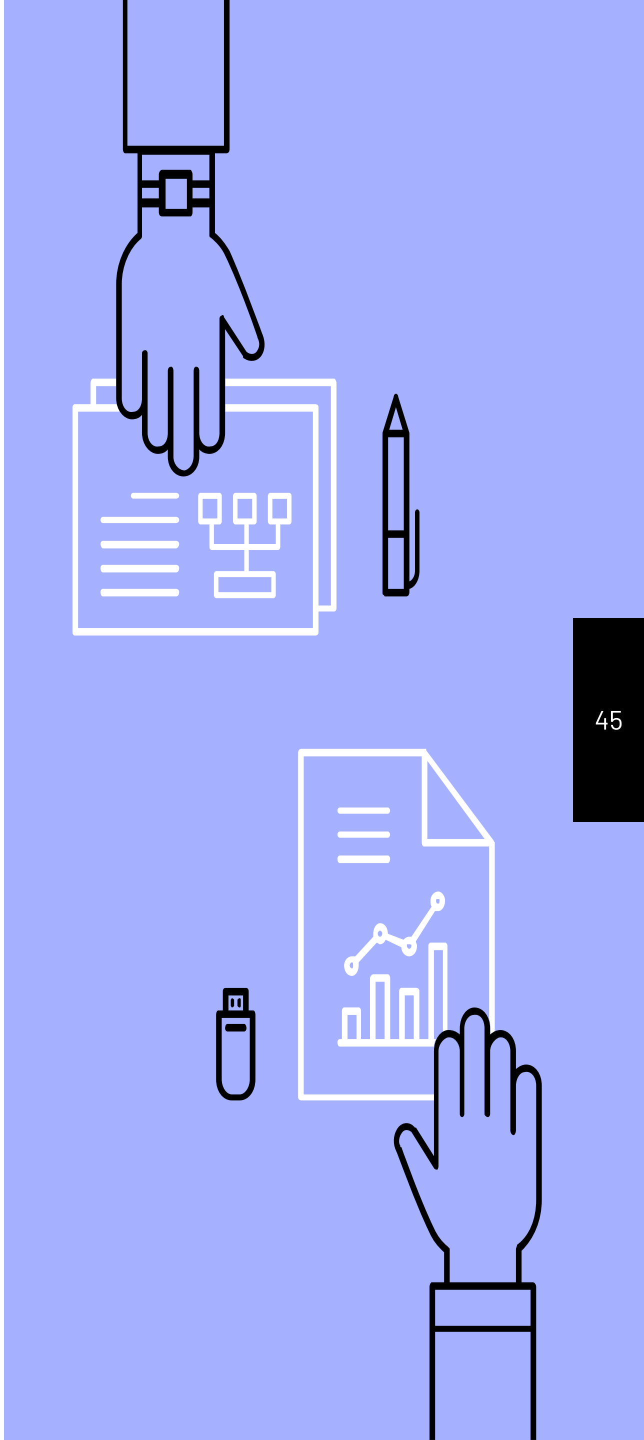
SAS example:

http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect060.htm

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash



Thank you

robertnakano@gmail.com
afsanehkhani35@gmail.com
joannewangziyi@gmail.com

Link to slides:
https://github.com/bobbyinfj/projects/tree/master/510%20Project-%20Categorical%20Response