Untitled

September 24, 2018

```
In [56]: import pandas as pd
         import numpy as np
         import seaborn as sns
         import matplotlib.pyplot as plt

In [57]: #Importing the data

In [58]: my_data_Shakespeare=pd.read_csv('C:/Users/Caio Laptop/OneDrive - The University of Ka

In [4]: #Taking a look in my data

In [8]: my_data_Shakespeare.head()

Out[8]:    Dataline      Play  PlayerLinenumber ActSceneLine         Player  \
        0         1  Henry IV               NaN          NaN            NaN
        1         2  Henry IV               NaN          NaN            NaN
        2         3  Henry IV               NaN          NaN            NaN
        3         4  Henry IV               1.0        1.1.1  KING HENRY IV
        4         5  Henry IV               1.0        1.1.2  KING HENRY IV

                                                PlayerLine
        0                                            ACT I
        1                        SCENE I. London. The palace.
        2  Enter KING HENRY, LORD JOHN OF LANCASTER, the ...
        3              So shaken as we are, so wan with care,
        4          Find we a time for frighted peace to pant,

In [7]: my_data_Shakespeare.shape

Out[7]: (111396, 6)

In [9]: my_data_Shakespeare.index

Out[9]: RangeIndex(start=0, stop=111396, step=1)

In [10]: my_data_Shakespeare.columns

Out[10]: Index(['Dataline', 'Play', 'PlayerLinenumber', 'ActSceneLine', 'Player',
                'PlayerLine'],
               dtype='object')
```

```
In [11]: my_data_Shakespeare.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 111396 entries, 0 to 111395
Data columns (total 6 columns):
Dataline            111396 non-null int64
Play                111396 non-null object
PlayerLinenumber    111393 non-null float64
ActSceneLine        105153 non-null object
Player              111389 non-null object
PlayerLine          111396 non-null object
dtypes: float64(1), int64(1), object(4)
memory usage: 3.4+ MB


In [12]: my_data_Shakespeare.count()

Out[12]: Dataline            111396
         Play                111396
         PlayerLinenumber    111393
         ActSceneLine        105153
         Player              111389
         PlayerLine          111396
         dtype: int64

In [13]: #Want to know how many Play we have

In [14]: my_data_Shakespeare["Play"].value_counts()

Out[14]: Hamlet                      4244
         Coriolanus                  3992
         Cymbeline                   3958
         Richard III                 3941
         Antony and Cleopatra        3862
         King Lear                   3766
         Othello                     3762
         Troilus and Cressida        3711
         A Winters Tale              3489
         Henry VIII                  3419
         Henry V                     3395
         Henry VI Part 2             3334
         Romeo and Juliet            3313
         Henry IV                    3205
         Henry VI Part 3             3138
         Alls well that ends well    3083
         Measure for measure         2998
         Loves Labours Lost          2986
         Henry VI Part 1             2983
         Richard II                  2937
```

```
Merry Wives of Windsor      2831
As you like it              2822
Taming of the Shrew         2806
Merchant of Venice          2802
Julius Caesar               2771
King John                   2766
Titus Andronicus            2726
Much Ado about nothing      2704
Timon of Athens             2662
Twelfth Night               2648
Pericles                    2641
macbeth                     2586
The Tempest                 2403
Two Gentlemen of Verona     2357
A Midsummer nights dream    2300
A Comedy of Errors          2055
Name: Play, dtype: int64
```

```
In [15]: #Want to know how many Player we have
```

```
In [16]: my_data_Shakespeare["Player"].value_counts()
```

```
Out[16]: GLOUCESTER          1920
         HAMLET              1582
         IAGO                1161
         FALSTAFF            1117
         KING HENRY V        1086
         BRUTUS              1051
         OTHELLO              928
         MARK ANTONY          927
         KING HENRY VI        917
         DUKE VINCENTIO       909
         TIMON                875
         QUEEN MARGARET       847
         Clown                804
         KING LEAR            801
         KING RICHARD II      794
         MACBETH              783
         TITUS ANDRONICUS     768
         PROSPERO             745
         CLEOPATRA            742
         YORK                 740
         HELENA               735
         LEONTES              720
         CORIOLANUS           717
         ROSALIND             711
         PORTIA               707
         WARWICK              690
```

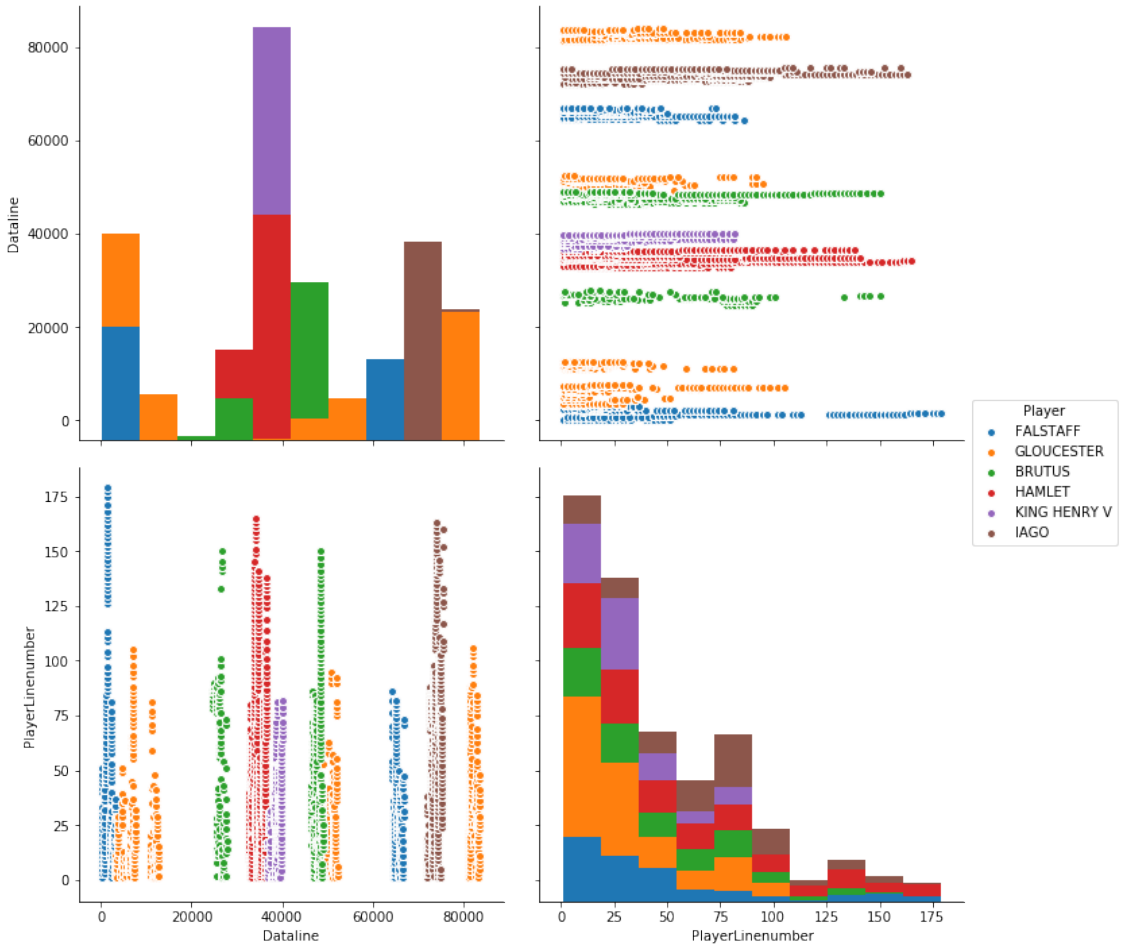
```
BUCKINGHAM           668
ROMEO                651
BIRON                647
PERICLES             645
                     ...
PHILEMON               1
ANOTHER                1
GURNEY                 1
HORTENSIA              1
A Lord                 1
Second Roman           1
Thieves                1
NICHOLAS               1
Haberdasher            1
Players                1
All Lords              1
Marshal                1
ARMADO                 1
JOSEPH                 1
Mariners               1
PHRYNIA                1
GLANSDALE              1
All The Lords          1
First Messenger        1
Ostler                 1
GENTLEMEN              1
Some Speak             1
Second Pirate          1
Third Musician         1
First Commoner         1
JOHN MORTIMER          1
MYRMIDONS              1
Musician               1
Soldiers               1
WORCESTER              1
Name: Player, Length: 934, dtype: int64
```

```
In [17]: #Creating new variables
```

```
In [18]: counts_player=my_data_Shakespeare["Player"].value_counts()
         counts_player[counts_player > 1000]
```

```
Out[18]: GLOUCESTER      1920
         HAMLET          1582
         IAGO            1161
         FALSTAFF        1117
         KING HENRY V    1086
         BRUTUS          1051
         Name: Player, dtype: int64
```

```
In [19]: counts_play=my_data_Shakespeare["Play"].value_counts()
         counts_play[counts_play > 3500]

Out[19]: Hamlet                 4244
         Coriolanus             3992
         Cymbeline              3958
         Richard III            3941
         Antony and Cleopatra   3862
         King Lear              3766
         Othello                3762
         Troilus and Cressida   3711
         Name: Play, dtype: int64

In [40]: # Let's work only with the top Players, i.e., those who appears more than 1000 times

In [34]: Player_top1000=my_data_Shakespeare[my_data_Shakespeare['Player'].isin(counts_player[c

In [ ]: #Taking a look in my new dataset: Player_top1000

In [35]: Player_top1000.head()

Out[35]:      Dataline      Play  PlayerLinenumber ActSceneLine    Player  \
         114       115  Henry IV               1.0        1.2.1  FALSTAFF
         126       127  Henry IV               3.0       1.2.13  FALSTAFF
         127       128  Henry IV               3.0       1.2.14  FALSTAFF
         128       129  Henry IV               3.0       1.2.15  FALSTAFF
         129       130  Henry IV               3.0       1.2.16  FALSTAFF

                                                 PlayerLine
         114              Now, Hal, what time of day is it, lad?
         126  Indeed, you come near me now, Hal, for we that...
         127  purses go by the moon and the seven stars, and...
         128  by Phoebus, he,'that wandering knight so fair...
         129  I prithee, sweet wag, when thou art king, as, God

In [36]: Player_top1000.shape

Out[36]: (7917, 6)

In [ ]: # Notice that, compared to the original dataset, where we have 111,396 observations, n

In [38]: #Classification Models

In [39]: sns.pairplot(Player_top1000.drop("Play", axis=1), hue="Player", size=5)

Out[39]: <seaborn.axisgrid.PairGrid at 0xa171c30>
```




```
In [ ]: # It's clear that based on the 'Dataline' we can determine the player fairly well

In [41]: # Now, let's work only with the top Plays, i.e., those who appears more than 3500 tim

In [42]: Play_top3500=my_data_Shakespeare[my_data_Shakespeare['Play'].isin(counts_play[counts_

In [43]: #Taking a look in my new dataset: Play_top3500

In [44]: Play_top3500.head()
```

Out[44]:

	Dataline	Play	PlayerLinenumber	ActSceneLine	\
18565	18566	Antony and Cleopatra	52.0	NaN	
18566	18567	Antony and Cleopatra	52.0	NaN	
18567	18568	Antony and Cleopatra	52.0	NaN	
18568	18569	Antony and Cleopatra	1.0	1.1.1	
18569	18570	Antony and Cleopatra	1.0	1.1.2	

Player	PlayerLine

```
18565  ROSALIND                                                ACT I
18566  ROSALIND  SCENE I. Alexandria. A room in CLEOPATRA's pal...
18567  ROSALIND                          Enter DEMETRIUS and PHILO
18568     PHILO              Nay, but this dotage of our general's
18569     PHILO      O'erflows the measure: those his goodly eyes,
```

In [45]:
```
Play_top3500.shape
```

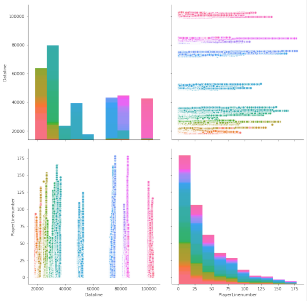
Out[45]:
```
(31236, 6)
```

In [ ]:
```
# Notice that, compared to the original dataset, where we have 111,396 observations, n
```

In [38]:
```
#Classification Models
```

In [47]:
```
sns.pairplot(Play_top3500.drop("Play", axis=1), hue="Player", size=5)
```

Out[47]:
```
<seaborn.axisgrid.PairGrid at 0x5d74e30>
```


Character
100000
80000
60000
40000
20000
Player Line number
125
100
75
50
25
0
20000
40000
60000
80000
100000
Character
0
25
50
75
100
125
150
175
Player Line number


Player
PERICLES
PHILO
CLEOPATRA
MARK ANTONY
Attendant
DEMETRIUS
CHARMIAN
ALEXAS
Soothsayer
DOMITIUS ENOBARBUS
IRAS
Messenger
First Attendant
Second Attendant
Second Messenger
OCTAVIUS CAESAR
LEPIDUS
MARDIAN
POMPEY
MENECRATES
MENAS
VARRIUS
TAURUS
AGRIPPA
OCTAVIA
Attendants
First Servant
Second Servant
VENTIDIUS
SILIUS
EROS
Soothsayers
Soldier
TAURUS
SCARUS
All
DOLABELLA
EUPHRONIUS
THYREUS
First Soldier
Second Soldier
Third Soldier
Fourth Soldier
Captain
First Guard
Second Guard
Third Guard
DERCETAS
DIOMEDES
Egyptian
PROCULEIUS
GALLUS
SELEUCUS
Clown
Queen
DROMIO OF EPHESUS
First Citizen
Second Citizen
MENENIUS
MARCIUS
First Senator
COMINIUS
TITUS
SICINIUS
BRUTUS
AUFIDIUS
Second Senator
VOLUMNIA
VIRGILIA
Gentlewoman
VALERIA
LARTIUS
First Roman
Second Roman
Third Roman
Lieutenant
CORIOLANUS
Both
Herald
First Officer
Second Officer
Officer
Tenants
Third Citizen
Fourth Citizen
Fifth Citizen
Both Citizens
Sixth Citizen
Seventh Citizen
All Citizens
Citizens
Aedile
A Patrician
Another Patrician
Both Tribunes
Roman
Volce
Citizen
First Servingman
Second Servingman
Third Servingman
Young MARCIUS
First Conspirator
Second Conspirator
Third Conspirator
All The Lords
Lords
First Lord
All Conspirators
All The People
Second Lord
Third Lord
First Gentleman
Second Gentleman
Messenger
POSTHUMUS LEONATUS
IMOGEN
CYMBELINE
PISANIO
CLOTEN
Lady
IACHIMO
PHILARIO
Frenchman
DUTCHMAN
CORNELIUS
CAIUS LUCIUS
BELARIUS
GUIDERIUS
ARVIRAGUS
First Tribune
Lord
First Captain
Second Captain
First Gaoler
Second Gaoler
Sicilius Leonatus
Mother
First Brother
Second Brother
Jupiter
Posthumus Leonatus
BERNARDO
FRANCISCO
HORATIO
MARCELLUS
KING CLAUDIUS
VOLTIMAND
LAERTES
LORD POLONIUS
HAMLET
QUEEN GERTRUDE
OPHELIA
Ghost
REYNALDO
ROSENCRANTZ
GUILDENSTERN
First Player
Prologue
Player King
Player Queen
LUCIANUS
PRINCE FORTINBRAS
Gentleman
Ravies
Sailor
First Sailor
First Clown
Second Clown
First Priest
OSRIC
First Ambassador
OCTAVIUS
KENT
GLOUCESTER
EDMUND
KING LEAR
GONERIL
CORDELIA
LEAR
REGAN
CORNWALL
BURGUNDY
KING OF FRANCE
EDGAR
OSWALD
Knight
Fool
ALBANY
CURAN
Third Servant
Old Man
Doctor
BENEDICK
RODERIGO
BIANCA
MONTANO
OTHELLO
CASSIO
DUKE OF VENICE
Sailor
Senator
DESDEMONA
MONTANO
Third Gentleman
Fourth Gentleman
Three Gentlemen
EMILIA
First Musician
BIANCA
LODOVICO
GRATIANO
HENRY BOLINGBROKE
CLARENCE
BRAKENBURY
HASTINGS
LADY ANNE
LORD MAYOR
RIVERS
GREY
QUEEN ELIZABETH
BUCKINGHAM
DERBY
QUEEN MARGARET
DORSET
CATESBY
First murderer
Second Murderer
Second murderer
KING EDWARD IV
Boy
DUCHESS OF YORK
Girl
Children
ARCHBISHOP OF YORK
YORK
PRINCE EDWARD
RICHARD
CARDINAL
STANLEY
LORD STANLEY
Pursuivant
Priest
RATCLIFF
LIEUTENANT
BISHOP OF ELY
LOVEL
Messenger
ANOTHER
KING RICHARD III
Page
TYRREL
Third Messenger
Fourth Messenger
CHRISTOPHER
Sheriff
RICHMOND
OXFORD
HERBERT
BLUNT
SURREY
NORFOLK
of Prince Edward
of King Henry VI
Ghost of CLARENCE
Ghost of RIVERS
Ghost of GREY
Ghost of VAUGHAN
Ghost of HASTINGS
of young Princes
Ghost of LADY ANNE
of BUCKINGHAM
LORDS
LUCIUS
TROILUS
PANDARUS
AENEAS
CRESSIDA
ALEXANDER
AGAMEMNON
NESTOR
ULYSSES
MENELAUS
AJAX
THERSITES
ACHILLES
PATROCLUS
PRIAM
HECUBA
HELENUS
CASSANDRA
PARIS
HELEN
CALCHAS
DEIPHOBUS
ALL
ANDROMACHE
MARGARELON
MYRMIDONS

In [48]:
```
# Finally, let's work with the top 6 players (those who appear more than 1000 times i
# the top 8 plays (those who appear more than 3500 times in our original dataset)

# Creating a new dataset
```

In [50]:
```
my_data_Shakespeare_new=my_data_Shakespeare[my_data_Shakespeare['Play'].isin(counts_pl
my_data_Shakespeare_new=my_data_Shakespeare_new[my_data_Shakespeare_new['Player'].isi
```

In [51]:
```
#Taking a look in my new dataset
```

In [52]:
```
my_data_Shakespeare_new.head()
```

Out[52]:

	Dataline	Play	PlayerLinenumber	ActSceneLine	Player	PlayerLine
24756	24757	Coriolanus	78.0	1.1.265	BRUTUS	He has no equal.
24758	24759	Coriolanus	80.0	1.1.267	BRUTUS	Mark'd you his lip and eyes?
24760	24761	Coriolanus	82.0	1.1.269	BRUTUS	Being moved, he will not spare to gird the gods.
24762	24763	Coriolanus	84.0	1.1.271	BRUTUS	The present wars devour him: he is grown
24763	24764	Coriolanus	84.0	1.1.272	BRUTUS	Too proud to be so valiant.

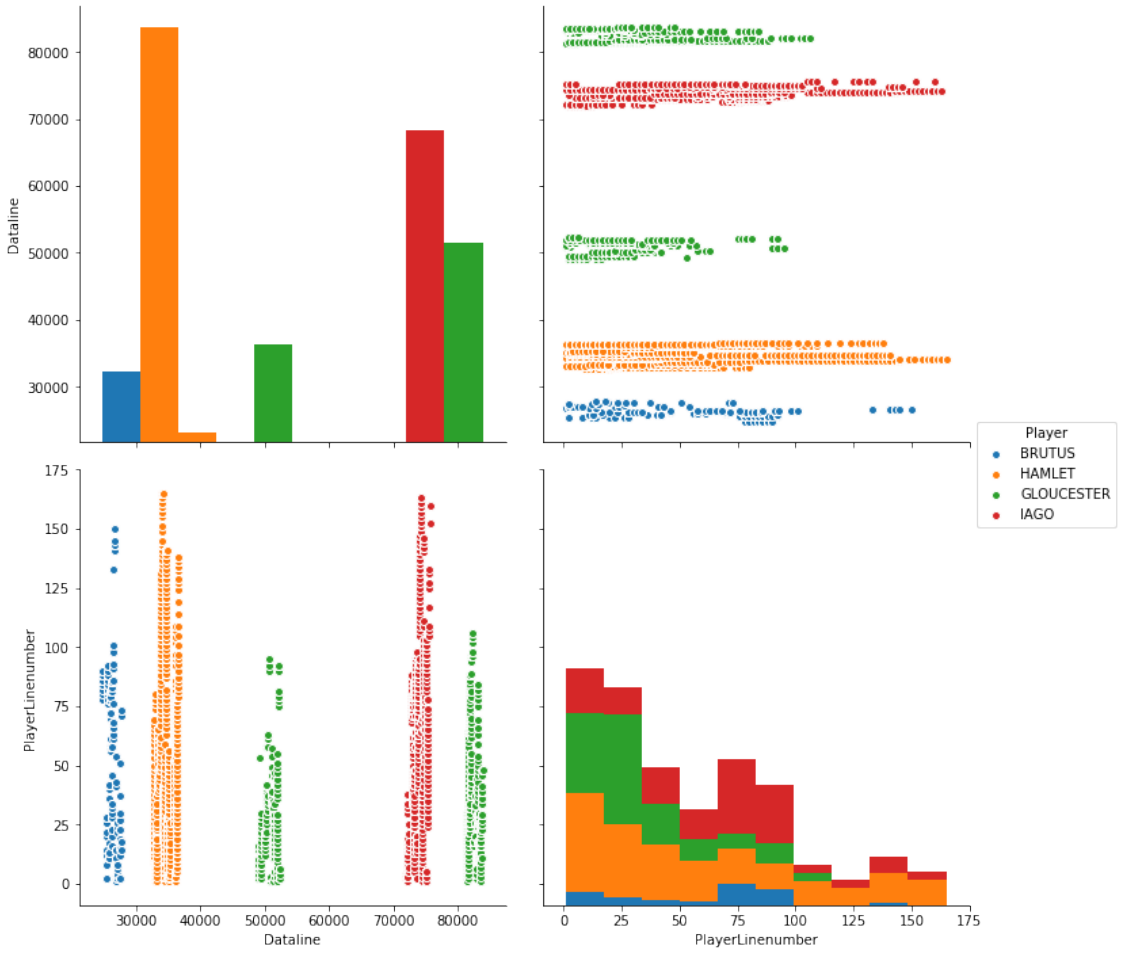
In [53]:
```
my_data_Shakespeare_new.shape
```

Out[53]:
```
(4112, 6)
```

In [54]:
```
# Notice that, compared to the original dataset, where we have 111,396 observations,
```

In [55]:
```
#Classification Models

sns.pairplot(my_data_Shakespeare_new.drop("Play", axis=1), hue="Player", size=5)
```

Out[55]:
```
<seaborn.axisgrid.PairGrid at 0xf291890>
```



In [ ]: *# It's clear that based on the 'Dataline' we can determine the player fairly well*