task

April 25, 2022

1 EDA

[ ]:
```
# importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[ ]:
```
# reading data and saving those into variables
ebrd = pd.read_csv("eclipse_bug_report_data.csv")
fbrd = pd.read_csv("freedesktop_bug_report_data.csv")
gcbrd = pd.read_csv("gcc_bug_report_data.csv")
gnbrd = pd.read_csv("gnome_bug_report_data.csv")
mbrd = pd.read_csv("mozilla_bug_report_data.csv")
sds = pd.read_csv("sales_data_sample.csv", encoding='Latin-1')
wbrd = pd.read_csv("winehq_bug_report_data.csv")
```

1.1 So we are going to perform EDA one by one as we have different CSV files

1.1.1 EDA On Eclipse_Bug_Report_Data

[ ]:
```
ebrd.head()
```

[ ]:
```
                    bug_id creation_date component_name       product_name  \
0       RECOMMENDERS-467951    2015-05-22           Core       RECOMMENDERS
1              QVTO-463396    2015-03-29         Engine               QVTO
2           EQUINOX-530069    2018-01-20     Compendium            EQUINOX
3          NATTABLE-422482    2013-11-25           Core           NATTABLE
4  WTP_JAVA_EE_TOOLS-116294    2005-11-14       jst.j2ee  WTP_JAVA_EE_TOOLS

                                  short_description  \
0  LogTraceException in ProposalUtils.toMethodNam…
1              CCE in DecorationNodeImpl.eSet (159)
2  [http servlet] During dispatching javax.servle…
3              Left border of NatTable is not drawn
4          NPE while importing EAR with utility jar

                                  long_description       assignee_name  \
```

```
0  The following incident was reported via the au…  recommenders-inbox
1  The following incident was reported via the au…      serg.boyko2011
2  Original issue https://issues.liferay.com/brow…        raymond.auge
3  Rendering a NatTable on a Composite with margi…          dirk.fauth
4  Import and EAR which has EJb and Ejb client tr…               jsholl

         reporter_name resolution_category  resolution_code status_category  \
0  error-reports-inbox               fixed               1          closed
1  error-reports-inbox               fixed               1        resolved
2         raymond.auge               fixed               1        resolved
3           dirk.fauth               fixed               1          closed
4             nagrawal               fixed               1          closed

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            6  2015-05-27                  0                     2
1            4  2015-04-01                  0                     8
2            4  2018-01-22                  0                     3
3            6  2014-07-23                  0                     3
4            6  2005-12-09                  0                     4

  resolution_date  bug_fix_time severity_category  severity_code
0      2015-05-27             5            normal              2
1      2015-03-31             2            normal              2
2      2018-01-22             2            normal              2
3      2013-11-25             0            normal              2
4      2005-11-15             1           blocker              6
```

[ ]:
```
ebrd.tail()
```

[ ]:
```
                        bug_id creation_date   component_name  \
9771                 CDT-36699    2003-04-21         cdt-core
9772             PLATFORM-72596    2004-08-25               UI
9773  WTP_SOURCE_EDITING-293504    2009-10-27          jst.jsp
9774          Z_ARCHIVED-219989    2008-02-22  TPTP.monitoring
9775                 PDE-143523    2006-05-24               UI

            product_name                                  short_description  \
9771                 CDT   Problem parsing Loki's Reference SmartPtr.h Impl
9772            PLATFORM  [Tests] Porting keyboard events from AWT's rob…
9773  WTP_SOURCE_EDITING  [validation] JSP syntax validator requires bra…
9774          Z_ARCHIVED  Update MAX test cases to reflect changes done …
9775                 PDE                           NPE during plugin-import

                                      long_description    assignee_name  \
9771  template\n    <\n        template  class Threa…         jcamelon
9772  In 3.0 SWT introduced a Display.post(Event) me…  douglas.pollock
9773  Problem also exists in 3.0.5p\n\n+++ This bug …         itewksbu
```

```
9774  Hari\n\nplease update the MAX manual and junit…         harihnar
9775  -Start a runtime workbench with branch32 pde c…   mike.pawlowski

     reporter_name resolution_category  resolution_code status_category  \
9771      jcamelon               fixed               1        resolved
9772          ines               fixed               1        resolved
9773      itewksbu               fixed               1        resolved
9774         apnan               fixed               1          closed
9775      janek.lb               fixed               1        resolved

      status_code update_date  quantity_of_votes  quantity_of_comments  \
9771            4  2003-04-27                  0                     3
9772            4  2004-08-27                  0                     3
9773            4  2010-01-22                  0                     6
9774            6  2010-06-03                  0                     6
9775            4  2006-06-22                  0                     4

     resolution_date  bug_fix_time severity_category  severity_code
9771      2003-04-27             6            normal              2
9772      2004-08-25             0            normal              2
9773      2010-01-22            87            normal              2
9774      2008-03-25            32            normal              2
9775      2006-05-24             0            normal              2
```

[ ]:
```
ebrd.describe()
```

[ ]:
```
       resolution_code  status_code  quantity_of_votes  quantity_of_comments  \
count           9776.0  9776.000000        9776.000000           9776.000000
mean               1.0     4.694763           0.048691              5.864669
std                0.0     0.952325           0.719535              5.825510
min                1.0     4.000000           0.000000              1.000000
25%                1.0     4.000000           0.000000              3.000000
50%                1.0     4.000000           0.000000              4.000000
75%                1.0     6.000000           0.000000              7.000000
max                1.0     6.000000          46.000000            182.000000

       bug_fix_time  severity_code
count   9776.000000    9776.000000
mean     102.394538       2.373261
std      302.926231       0.965967
min        0.000000       1.000000
25%        1.000000       2.000000
50%        8.000000       2.000000
75%       60.000000       2.000000
max     4961.000000       6.000000
```

[ ]:
```
ebrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9776 entries, 0 to 9775
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                9776 non-null   object
 1   creation_date         9776 non-null   object
 2   component_name        9776 non-null   object
 3   product_name          9776 non-null   object
 4   short_description     9776 non-null   object
 5   long_description      9627 non-null   object
 6   assignee_name         9776 non-null   object
 7   reporter_name         9776 non-null   object
 8   resolution_category   9776 non-null   object
 9   resolution_code       9776 non-null   int64
 10  status_category       9776 non-null   object
 11  status_code           9776 non-null   int64
 12  update_date           9776 non-null   object
 13  quantity_of_votes     9776 non-null   int64
 14  quantity_of_comments  9776 non-null   int64
 15  resolution_date       9776 non-null   object
 16  bug_fix_time          9776 non-null   int64
 17  severity_category     9776 non-null   object
 18  severity_code         9776 non-null   int64
dtypes: int64(6), object(13)
memory usage: 1.4+ MB
```

[ ]:
```
ebrd['status_category'].unique()
```

[ ]:
```
array(['closed', 'resolved'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status_category we can encode it as it is two categories only
# We are encoding it
ebrd['status_category'] = ebrd['status_category'].replace("resolved", 1)
ebrd['status_category'] = ebrd['status_category'].replace("closed", 0)
```

[ ]:
```
ebrd['severity_category'].unique()
```

[ ]:
```
array(['normal', 'blocker', 'trivial', 'minor', 'major', 'critical'],
      dtype=object)
```

[ ]:
```
# As we have another column severity_category and we can perform encoding on it␣
 ↪as well
ebrd['severity_category'] = ebrd['severity_category'].replace("normal", 0)
ebrd['severity_category'] = ebrd['severity_category'].replace("blocker", 1)
ebrd['severity_category'] = ebrd['severity_category'].replace("trivial", 2)
```

```
ebrd['severity_category'] = ebrd['severity_category'].replace("minor", 3)
ebrd['severity_category'] = ebrd['severity_category'].replace("major", 4)
ebrd['severity_category'] = ebrd['severity_category'].replace("critical", 5)
```

[ ]:
```
ebrd.head(15)
```

[ ]:

	bug_id	creation_date	component_name	\
0	RECOMMENDERS-467951	2015-05-22	Core	
1	QVTO-463396	2015-03-29	Engine	
2	EQUINOX-530069	2018-01-20	Compendium	
3	NATTABLE-422482	2013-11-25	Core	
4	WTP_JAVA_EE_TOOLS-116294	2005-11-14	jst.j2ee	
5	PDT-165702	2006-11-23	Code Assist	
6	WTP_SOURCE_EDITING-103608	2005-07-13	wst.html	
7	JDT-84477	2005-02-04	UI	
8	BIRT-236012	2008-06-06	Report Engine	
9	SUBVERSIVE-264685	2009-02-12	UI	
10	PDE-136135	2006-04-11	UI	
11	COMMUNITY-500202	2016-08-24	Marketplace	
12	PDT-310759	2010-04-28	Code Assist	
13	ECF-193415	2007-06-19	ecf.ui	
14	COMMUNITY-257135	2008-12-01	CVS	

	product_name	short_description	\
0	RECOMMENDERS	LogTraceException in ProposalUtils.toMethodNam…	
1	QVTO	CCE in DecorationNodeImpl.eSet (159)	
2	EQUINOX	[http servlet] During dispatching javax.servle…	
3	NATTABLE	Left border of NatTable is not drawn	
4	WTP_JAVA_EE_TOOLS	NPE while importing EAR with utility jar	
5	PDT	No Items in Outline/PHP Project Outline (Repro…	
6	WTP_SOURCE_EDITING	Error in setting disabled attribute	
7	JDT	[refactoring] Lose comments when refactoring	
8	BIRT	[Regression] There are two extra lines when pr…	
9	SUBVERSIVE	Share Project Wizard dialog typo	
10	PDE	[Schema][Editors] StackOverflowError in schema…	
11	COMMUNITY	Wrong username in message when user favorite i…	
12	PDT	No content assist in JavaScript regions	
13	ECF	[UI] Contacts view menu entries should be disa…	
14	COMMUNITY	Please create new directory under /cvsroot/ecl…	

	long_description	assignee_name	\
0	The following incident was reported via the au…	recommenders-inbox	
1	The following incident was reported via the au…	serg.boyko2011	
2	Original issue https://issues.liferay.com/brow…	raymond.auge	
3	Rendering a NatTable on a Composite with margi…	dirk.fauth	
4	Import and EAR which has EJb and Ejb client tr…	jsholl	
5	Using\n\norg.eclipse.php_feature (0.7.0.v20061…	guy.g	

```
6   1. Open an HTML file\n2. Type the  tag.\n3. Se…              nitind
7   build i0202\n\nHere is a test case in which yo…        jdt-ui-inbox
8   Created attachment 103916\nreport design\n\nDe…              hustlg
9   Created attachment 125521\nscreen shot of dial…         igor.burilo
10  N20060411-0010\n\nHappened on Definition tab a…      mike.pawlowski
11  Created attachment 263755\nScreen capture of a…   marketplace-inbox
12  No content assist in JavaScript regions ? sinc…      php.core-inbox
13  If you aren't logged in you can press the chev…      ecf.core-inbox
14  We would like to split off a new plug-in see b…           webmaster

          reporter_name resolution_category  resolution_code  status_category  \
0   error-reports-inbox               fixed                1                0
1   error-reports-inbox               fixed                1                1
2          raymond.auge               fixed                1                1
3            dirk.fauth               fixed                1                0
4              nagrawal               fixed                1                0
5                mickey               fixed                1                0
6            kvenugopal               fixed                1                1
7           dj.houghton               fixed                1                1
8                 xwang               fixed                1                0
9             mikecepek               fixed                1                1
10        markus.kell.r               fixed                1                1
11       antoine.thomas               fixed                1                1
12         zhaozhongwei               fixed                1                0
13           caniszczyk               fixed                1                0
14             bokowski               fixed                1                1

    status_code update_date  quantity_of_votes  quantity_of_comments  \
0             6  2015-05-27                  0                     2
1             4  2015-04-01                  0                     8
2             4  2018-01-22                  0                     3
3             6  2014-07-23                  0                     3
4             6  2005-12-09                  0                     4
5             6  2006-12-03                  0                     5
6             4  2007-08-29                  0                     3
7             4  2005-02-07                  0                     2
8             6  2009-09-08                  0                     4
9             4  2009-10-05                  0                     2
10            4  2007-05-11                  0                    13
11            4  2016-08-24                  0                     3
12            6  2010-05-27                  0                    12
13            6  2008-05-18                  0                     4
14            4  2008-12-01                  0                     2

   resolution_date  bug_fix_time  severity_category  severity_code
0       2015-05-27             5                  0              2
1       2015-03-31             2                  0              2
```

```
2       2018-01-22             2                  0              2
3       2013-11-25             0                  0              2
4       2005-11-15             1                  1              6
5       2006-12-03            10                  0              2
6       2007-08-29           777                  2              1
7       2005-02-07             3                  0              2
8       2008-06-08             2                  0              2
9       2009-10-05           235                  3              2
10      2007-05-11           395                  4              4
11      2016-08-24             0                  0              2
12      2010-05-18            20                  0              2
13      2007-06-20             1                  0              2
14      2008-12-01             0                  0              2
```

[ ]:
```
ebrd['component_name'].unique
```

[ ]:
```
<bound method Series.unique of 0                 Core
1                Engine
2            Compendium
3                  Core
4              jst.j2ee
             …
9771           cdt-core
9772                 UI
9773            jst.jsp
9774    TPTP.monitoring
9775                 UI
Name: component_name, Length: 9776, dtype: object>
```

[ ]:
```
ebrd['product_name'].unique
```

[ ]:
```
<bound method Series.unique of 0             RECOMMENDERS
1                    QVTO
2                 EQUINOX
3                NATTABLE
4        WTP_JAVA_EE_TOOLS
               …
9771                   CDT
9772              PLATFORM
9773    WTP_SOURCE_EDITING
9774            Z_ARCHIVED
9775                   PDE
Name: product_name, Length: 9776, dtype: object>
```

1.2 As other elements are not having a categorical

1.2.1 Now we will be removing null values

[ ]:
```
ebrd.isnull().sum()
```

[ ]:
```
bug_id                   0
creation_date            0
component_name           0
product_name             0
short_description        0
long_description       149
assignee_name            0
reporter_name            0
resolution_category      0
resolution_code          0
status_category          0
status_code              0
update_date              0
quantity_of_votes        0
quantity_of_comments     0
resolution_date          0
bug_fix_time             0
severity_category        0
severity_code            0
dtype: int64
```

[ ]:
```
# percentage of missing values
ebrd.isnull().sum() / ebrd.shape[0] * 100
```

[ ]:
```
bug_id                  0.000000
creation_date           0.000000
component_name          0.000000
product_name            0.000000
short_description       0.000000
long_description        1.524141
assignee_name           0.000000
reporter_name           0.000000
resolution_category     0.000000
resolution_code         0.000000
status_category         0.000000
status_code             0.000000
update_date             0.000000
quantity_of_votes       0.000000
quantity_of_comments    0.000000
resolution_date         0.000000
bug_fix_time            0.000000
severity_category       0.000000
```

```
severity_code           0.000000
dtype: float64
```

```
[ ]: ebrd.dropna(inplace=True)
```

```
[ ]: ebrd.isnull().sum()
```

```
[ ]: bug_id                 0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description        0
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

```
[ ]: ebrd.head()
```

```
[ ]:                    bug_id creation_date component_name       product_name  \
0       RECOMMENDERS-467951    2015-05-22           Core       RECOMMENDERS
1               QVTO-463396    2015-03-29         Engine               QVTO
2            EQUINOX-530069    2018-01-20     Compendium            EQUINOX
3          NATTABLE-422482    2013-11-25           Core           NATTABLE
4  WTP_JAVA_EE_TOOLS-116294    2005-11-14       jst.j2ee  WTP_JAVA_EE_TOOLS

                                 short_description  \
0  LogTraceException in ProposalUtils.toMethodNam…
1              CCE in DecorationNodeImpl.eSet (159)
2  [http servlet] During dispatching javax.servle…
3              Left border of NatTable is not drawn
4          NPE while importing EAR with utility jar

                                  long_description       assignee_name  \
0  The following incident was reported via the au…  recommenders-inbox
1  The following incident was reported via the au…      serg.boyko2011
```

```
2  Original issue https://issues.liferay.com/brow…        raymond.auge
3  Rendering a NatTable on a Composite with margi…          dirk.fauth
4  Import and EAR which has EJb and Ejb client tr…              jsholl

         reporter_name resolution_category  resolution_code  status_category  \
0  error-reports-inbox               fixed                1                0
1  error-reports-inbox               fixed                1                1
2         raymond.auge               fixed                1                1
3           dirk.fauth               fixed                1                0
4             nagrawal               fixed                1                0

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            6  2015-05-27                  0                     2
1            4  2015-04-01                  0                     8
2            4  2018-01-22                  0                     3
3            6  2014-07-23                  0                     3
4            6  2005-12-09                  0                     4

  resolution_date  bug_fix_time  severity_category  severity_code
0      2015-05-27             5                  0              2
1      2015-03-31             2                  0              2
2      2018-01-22             2                  0              2
3      2013-11-25             0                  0              2
4      2005-11-15             1                  1              6
```

[ ]:
```
from datetime import time,date,datetime

ebrd['creation_date'] = pd.to_datetime(ebrd['creation_date'])
ebrd['update_date'] = pd.to_datetime(ebrd['update_date'])
ebrd['resolution_date'] = pd.to_datetime(ebrd['resolution_date'])
```

[ ]:
```
ebrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9627 entries, 0 to 9775
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                9627 non-null   object
 1   creation_date         9627 non-null   datetime64[ns]
 2   component_name        9627 non-null   object
 3   product_name          9627 non-null   object
 4   short_description     9627 non-null   object
 5   long_description      9627 non-null   object
 6   assignee_name         9627 non-null   object
 7   reporter_name         9627 non-null   object
 8   resolution_category   9627 non-null   object
```

```
 9   resolution_code       9627 non-null   int64
 10  status_category       9627 non-null   int64
 11  status_code           9627 non-null   int64
 12  update_date           9627 non-null   datetime64[ns]
 13  quantity_of_votes     9627 non-null   int64
 14  quantity_of_comments  9627 non-null   int64
 15  resolution_date       9627 non-null   datetime64[ns]
 16  bug_fix_time          9627 non-null   int64
 17  severity_category     9627 non-null   int64
 18  severity_code         9627 non-null   int64
dtypes: datetime64[ns](3), int64(8), object(8)
memory usage: 1.5+ MB
```

[ ]:
```
ebrd.insert(0,"time_to_dep[s]",␣
↪((ebrd['resolution_date']-ebrd['creation_date']).astype('timedelta64[s]')),␣
↪True)
```

[ ]:
```
ebrd.head()
```

[ ]:
```
   time_to_dep[s]                   bug_id creation_date component_name  \
0        432000.0       RECOMMENDERS-467951    2015-05-22           Core
1        172800.0               QVTO-463396    2015-03-29         Engine
2        172800.0            EQUINOX-530069    2018-01-20     Compendium
3             0.0           NATTABLE-422482    2013-11-25           Core
4         86400.0  WTP_JAVA_EE_TOOLS-116294    2005-11-14       jst.j2ee

        product_name                                  short_description  \
0       RECOMMENDERS  LogTraceException in ProposalUtils.toMethodNam…
1               QVTO               CCE in DecorationNodeImpl.eSet (159)
2            EQUINOX  [http servlet] During dispatching javax.servle…
3           NATTABLE               Left border of NatTable is not drawn
4  WTP_JAVA_EE_TOOLS           NPE while importing EAR with utility jar

                                long_description       assignee_name  \
0  The following incident was reported via the au…  recommenders-inbox
1  The following incident was reported via the au…      serg.boyko2011
2  Original issue https://issues.liferay.com/brow…        raymond.auge
3  Rendering a NatTable on a Composite with margi…          dirk.fauth
4  Import and EAR which has EJb and Ejb client tr…              jsholl

         reporter_name resolution_category  resolution_code  status_category  \
0  error-reports-inbox              fixed                1                0
1  error-reports-inbox              fixed                1                1
2        raymond.auge               fixed                1                1
3          dirk.fauth               fixed                1                0
4            nagrawal               fixed                1                0
```

```
   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            6  2015-05-27                  0                     2
1            4  2015-04-01                  0                     8
2            4  2018-01-22                  0                     3
3            6  2014-07-23                  0                     3
4            6  2005-12-09                  0                     4

  resolution_date  bug_fix_time  severity_category  severity_code
0      2015-05-27             5                  0              2
1      2015-03-31             2                  0              2
2      2018-01-22             2                  0              2
3      2013-11-25             0                  0              2
4      2005-11-15             1                  1              6
```

[ ]:
```
# dropping extra columns
ebrd.
 ↪drop(['bug_id','resolution_code','creation_date','component_name','product_name','short_des
```

[ ]:
```
# now we are done with dropping values as well
ebrd['bug_fix_time'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew         6.535351
kurtosis    59.143279
Name: bug_fix_time, dtype: float64
```

[ ]:
```
corr = ebrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.038578    -0.038578
status_category             0.038578         1.000000    -1.000000
status_code                -0.038578        -1.000000     1.000000
quantity_of_votes           0.131352         0.026819    -0.026819
quantity_of_comments        0.134959        -0.025127     0.025127
bug_fix_time                1.000000         0.038578    -0.038578
severity_category          -0.019772        -0.094191     0.094191
severity_code              -0.048731        -0.112924     0.112924

                      quantity_of_votes  quantity_of_comments  bug_fix_time  \
time_to_dep[s]                 0.131352              0.134959      1.000000
status_category                0.026819             -0.025127      0.038578
status_code                   -0.026819              0.025127     -0.038578
quantity_of_votes              1.000000              0.362044      0.131352
quantity_of_comments           0.362044              1.000000      0.134959
bug_fix_time                   0.131352              0.134959      1.000000
severity_category              0.025544              0.109243     -0.019772
severity_code                  0.031407              0.135848     -0.048731
```

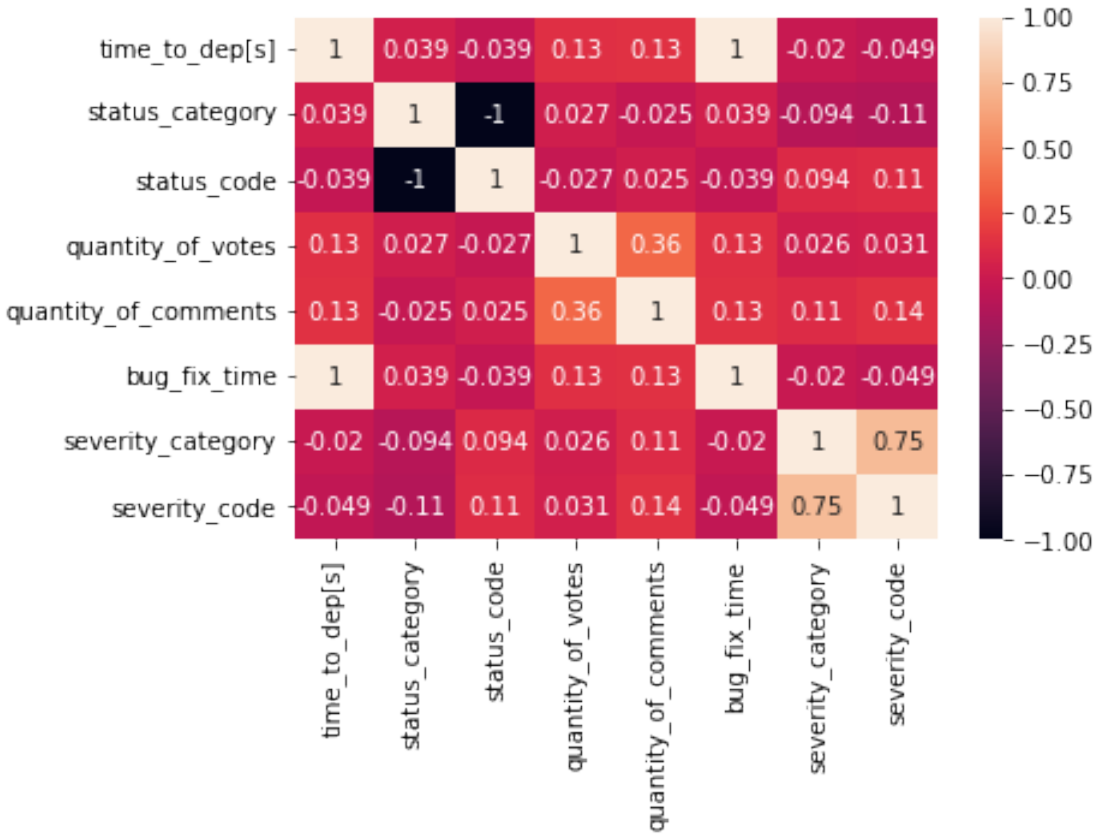
```
                        severity_category  severity_code
time_to_dep[s]                  -0.019772      -0.048731
status_category                 -0.094191      -0.112924
status_code                      0.094191       0.112924
quantity_of_votes                0.025544       0.031407
quantity_of_comments             0.109243       0.135848
bug_fix_time                    -0.019772      -0.048731
severity_category                1.000000       0.745366
severity_code                    0.745366       1.000000
```

[ ]:
```
corr = ebrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                        time_to_dep[s]  status_category  status_code  \
time_to_dep[s]                1.000000         0.038578    -0.038578
status_category               0.038578         1.000000    -1.000000
status_code                  -0.038578        -1.000000     1.000000
quantity_of_votes             0.131352         0.026819    -0.026819
quantity_of_comments          0.134959        -0.025127     0.025127
bug_fix_time                  1.000000         0.038578    -0.038578
severity_category            -0.019772        -0.094191     0.094191
severity_code                -0.048731        -0.112924     0.112924

                        quantity_of_votes  quantity_of_comments  bug_fix_time  \
time_to_dep[s]                   0.131352              0.134959      1.000000
status_category                  0.026819             -0.025127      0.038578
status_code                     -0.026819              0.025127     -0.038578
quantity_of_votes                1.000000              0.362044      0.131352
quantity_of_comments             0.362044              1.000000      0.134959
bug_fix_time                     0.131352              0.134959      1.000000
severity_category                0.025544              0.109243     -0.019772
severity_code                    0.031407              0.135848     -0.048731

                        severity_category  severity_code
time_to_dep[s]                  -0.019772      -0.048731
status_category                 -0.094191      -0.112924
status_code                      0.094191       0.112924
quantity_of_votes                0.025544       0.031407
quantity_of_comments             0.109243       0.135848
bug_fix_time                    -0.019772      -0.048731
severity_category                1.000000       0.745366
severity_code                    0.745366       1.000000
```

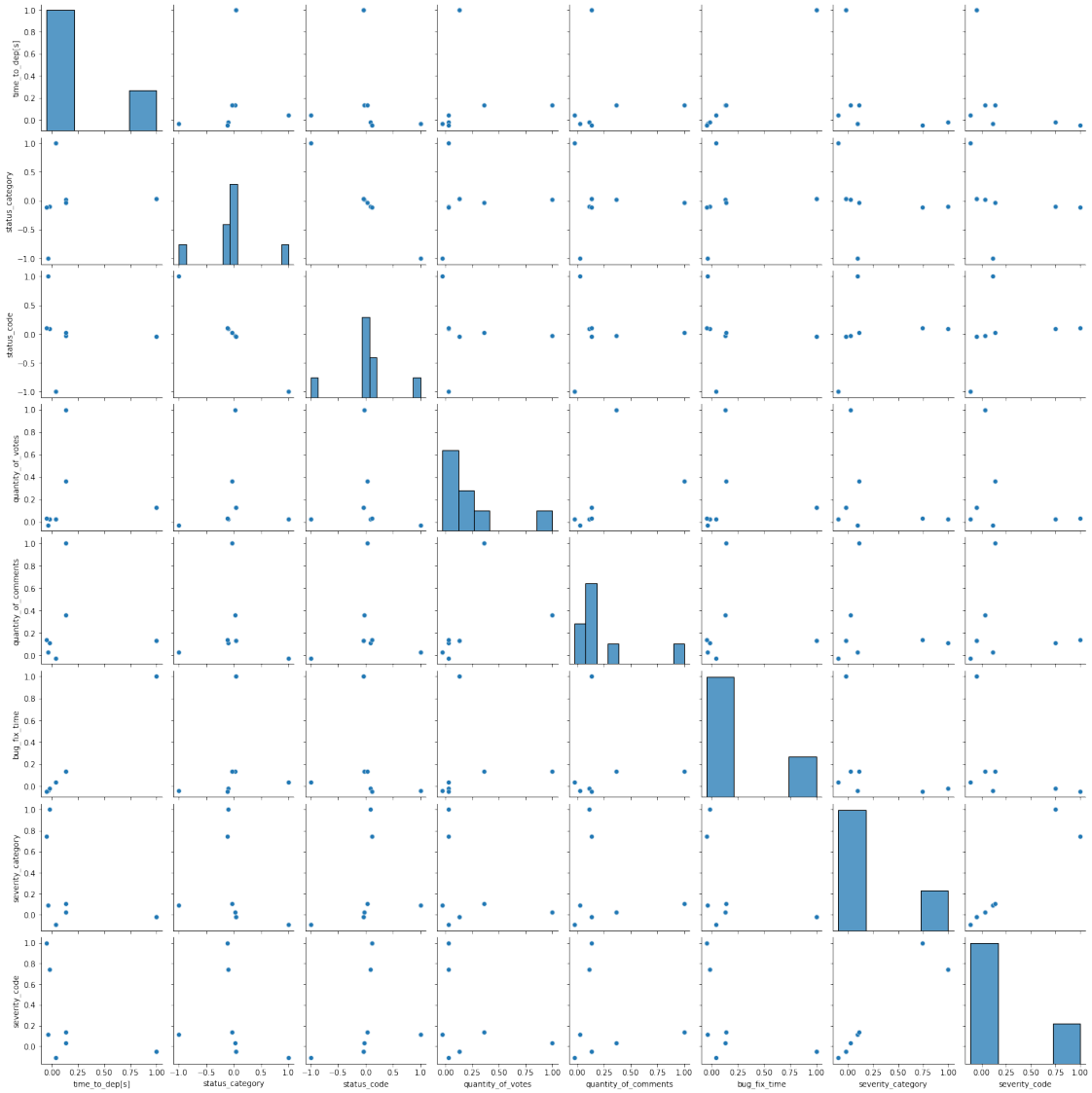
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]:
```
<AxesSubplot:>
```



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]:
```
<seaborn.axisgrid.PairGrid at 0x266900b8700>
```



[ ]:
```
ebrd.head()
```

[ ]:

	time_to_dep[s]	status_category	status_code	quantity_of_votes	quantity_of_comments	bug_fix_time	severity_category	severity_code
0	432000.0	0	6	0	2	5	0	2
1	172800.0	1	4	0	8	2	0	2
2	172800.0	1	4	0	3	2	0	2
3	0.0	0	6	0				
4	86400.0	0	6	0				

```
3                    3              0                    0              2
4                    4              1                    1              6
```

[ ]:
```
X = ebrd.iloc[:,:-1].values #rows and then columns in brackets
Y = ebrd.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[4.3200e+05, 0.0000e+00, 6.0000e+00, …, 2.0000e+00, 5.0000e+00,
        0.0000e+00],
       [1.7280e+05, 1.0000e+00, 4.0000e+00, …, 8.0000e+00, 2.0000e+00,
        0.0000e+00],
       [1.7280e+05, 1.0000e+00, 4.0000e+00, …, 3.0000e+00, 2.0000e+00,
        0.0000e+00],
       …,
       [7.5168e+06, 1.0000e+00, 4.0000e+00, …, 6.0000e+00, 8.7000e+01,
        0.0000e+00],
       [2.7648e+06, 0.0000e+00, 6.0000e+00, …, 6.0000e+00, 3.2000e+01,
        0.0000e+00],
       [0.0000e+00, 1.0000e+00, 4.0000e+00, …, 4.0000e+00, 0.0000e+00,
        0.0000e+00]])
```

[ ]:
```
Y
```

[ ]:
```
array([2, 2, 2, …, 2, 2, 2], dtype=int64)
```

1.2.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
#
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=42)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
        # assignment which one we should accept from these
```

```
LinearRegression()
R2 score is:  0.5497211472623571
Mean Absolute error is:  0.23423773171339624
Mean Squared Error:  0.40367775495542163
---------------------------------------------------
LogisticRegression()
R2 score is:  -0.14903198491205116
Mean Absolute error is:  0.3956386292834891
Mean Squared Error:  1.0301142263759087
---------------------------------------------------
GaussianNB()
R2 score is:  -0.14903198491205116
Mean Absolute error is:  0.3956386292834891
Mean Squared Error:  1.0301142263759087
---------------------------------------------------
RandomForestClassifier()
R2 score is:  0.9994208508140564
Mean Absolute error is:  0.0005192107995846313
Mean Squared Error:  0.0005192107995846313
---------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
---------------------------------------------------
SVR()
R2 score is:  -0.07848473467001504
Mean Absolute error is:  0.4648861541234049
Mean Squared Error:  0.9668681835674615
---------------------------------------------------
```

1.2.3 EDA On Freedesktop_bug_report_data

[ ]:
```
fbrd.head()
```

[ ]:
```
                  bug_id creation_date component_name     product_name  \
0  FREEDESKTOP.ORG-24230    2009-09-30   New Accounts  FREEDESKTOP.ORG
1              XORG-2056    2004-12-12  Documentation             XORG
2             XORG-79994    2014-06-13   Driver/intel             XORG
3     LIBREOFFICE-45271    2012-01-26         Writer      LIBREOFFICE
4       POLICYKIT-83093    2014-08-26         daemon        POLICYKIT

                                 short_description  \
0  New account request for mesa/r300/r600 develop…
1  XChangeProperty prototype missing from manual …
2  intel-virtual-output: fatal IO error 9 when ex…
3  : The way Writer counting Chinese words is not…
4       [patch] pkexec parameter parsing memory leak

                                  long_description  assignee_name  \
0  Created attachment 29953\nssh-pub\n\nWould lik…  sitewranglers
1  The manual page for XChangeProperty lists seve…      xorg-team
2  Created attachment 100992\ni-v-o debug output\…          chris
3  Problem description: \n\nThe way of counting C…        caolanm
4  Google's Project Zero has developed an exploit…        zeuthen

    reporter_name resolution_category  resolution_code status_category  \
0       amaasikas               fixed               1        resolved
1  matthieu.herrb               fixed               1        resolved
2      main.haarp               fixed               1        resolved
3        blackjay               fixed               1          closed
4           hanno               fixed               1        resolved

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            4  2009-09-30                  0                     4
1            4  2004-12-14                  0                     5
2            4  2014-06-13                  0                    10
3            6  2012-04-05                  0                     4
4            4  2014-08-27                  0                     6

  resolution_date  bug_fix_time severity_category  severity_code
0      2009-09-30             0            normal              2
1      2004-12-14             2            normal              2
2      2014-06-13             0            normal              2
3      2012-04-05            70             major              4
4      2014-08-27             1            normal              2
```

[ ]:
```
fbrd.tail()
```

[ ]:
```
                     bug_id creation_date component_name       product_name  \
7679           WAYLAND-93836    2016-01-23         weston            WAYLAND
7680              XORG-47203    2012-03-11         Lib/Xt               XORG
7681              XORG-21626    2009-05-07   Driver/intel               XORG
7682        PULSEAUDIO-66774    2013-07-10           core          PULSEAUDIO
7683  MEDIA-PLAYER-INFO-30725    2010-10-09        General  MEDIA-PLAYER-INFO

                                    short_description  \
7679  Weston crashes after clicking a Firefox menu w…
7680                    libxt 1.1.2 breaks xscreensaver
7681  [i945 DRI1] Cursor movement screenshot clippin…
7682  crash when shutting down without any modules l…
7683           Don't need to specify OS in device names

                                     long_description    assignee_name  \
7679  Created attachment 121236\ngdb backtrace\n\nHo…     wayland-bugs
7680  libxt 1.1.2 breaks some themes in xscreensaver…        xorg-team
7681  Created attachment 25621\nXorg.0.log\n\nForwar…          jbarnes
7682  If all the modules given fail to load the daem…  pulseaudio-bugs
7683  Created attachment 39298\npatch\n\nThe attache…      martin.pitt

        reporter_name resolution_category  resolution_code status_category  \
7679             bugs               fixed               1        resolved
7680   davemorgan353               fixed               1        resolved
7681            bryce               fixed               1        resolved
7682  pierre-bugzilla               fixed               1        resolved
7683         jonathan               fixed               1        resolved

      status_code update_date  quantity_of_votes  quantity_of_comments  \
7679            4  2017-04-06                  0                     3
7680            4  2012-03-12                  0                     8
7681            4  2009-10-08                  0                    24
7682            4  2017-04-27                  0                     2
7683            4  2010-10-09                  0                     2

     resolution_date  bug_fix_time severity_category  severity_code
7679      2017-04-06           439             major              4
7680      2012-03-12             1             major              4
7681      2009-10-05           151             major              4
7682      2013-07-11             1            normal              2
7683      2010-10-09             0            normal              2
```

[ ]:
```
fbrd.describe()
```

[ ]:
```
        resolution_code  status_code  quantity_of_votes  quantity_of_comments  \
count            7684.0  7684.000000             7684.0           7684.000000
mean                1.0     4.252993                0.0              8.355023
```

```
std                0.0     0.664860                0.0              10.340132
min                1.0     4.000000                0.0               1.000000
25%                1.0     4.000000                0.0               3.000000
50%                1.0     4.000000                0.0               5.000000
75%                1.0     4.000000                0.0              10.000000
max                1.0     6.000000                0.0             310.000000

       bug_fix_time  severity_code
count    7684.00000    7684.000000
mean      173.53670       2.486205
std       385.12708       1.094015
min         0.00000       0.000000
25%         3.00000       2.000000
50%        28.00000       2.000000
75%       162.00000       2.000000
max      4896.00000       6.000000
```

[ ]:
```
fbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7684 entries, 0 to 7683
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                7684 non-null   object
 1   creation_date         7684 non-null   object
 2   component_name        7684 non-null   object
 3   product_name          7684 non-null   object
 4   short_description     7684 non-null   object
 5   long_description      7636 non-null   object
 6   assignee_name         7684 non-null   object
 7   reporter_name         7684 non-null   object
 8   resolution_category   7684 non-null   object
 9   resolution_code       7684 non-null   int64
 10  status_category       7684 non-null   object
 11  status_code           7684 non-null   int64
 12  update_date           7684 non-null   object
 13  quantity_of_votes     7684 non-null   int64
 14  quantity_of_comments  7684 non-null   int64
 15  resolution_date       7684 non-null   object
 16  bug_fix_time          7684 non-null   int64
 17  severity_category     7684 non-null   object
 18  severity_code         7684 non-null   int64
dtypes: int64(6), object(13)
memory usage: 1.1+ MB
```

[ ]:
```
fbrd['status_category'].unique()
```

[ ]:
```
array(['resolved', 'closed'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status_category we can encode it as it is two categories only
# We are encoding it
fbrd['status_category'] = fbrd['status_category'].replace("resolved", 1)
fbrd['status_category'] = fbrd['status_category'].replace("closed", 0)
```

[ ]:
```
fbrd['severity_category'].unique()
```

[ ]:
```
array(['normal', 'major', 'critical', 'blocker', 'minor', 'trivial',
       'not set'], dtype=object)
```

[ ]:
```
# As we have another column severity_category and we can perform encoding on it␣
 ↪as well
fbrd['severity_category'] = fbrd['severity_category'].replace("normal", 0)
fbrd['severity_category'] = fbrd['severity_category'].replace("blocker", 1)
fbrd['severity_category'] = fbrd['severity_category'].replace("trivial", 2)
fbrd['severity_category'] = fbrd['severity_category'].replace("minor", 3)
fbrd['severity_category'] = fbrd['severity_category'].replace("major", 4)
fbrd['severity_category'] = fbrd['severity_category'].replace("critical", 5)
fbrd['severity_category'] = fbrd['severity_category'].replace("not set", 6)
```

[ ]:
```
fbrd.head(15)
```

[ ]:

	bug_id	creation_date	component_name	\
0	FREEDESKTOP.ORG-24230	2009-09-30	New Accounts	
1	XORG-2056	2004-12-12	Documentation	
2	XORG-79994	2014-06-13	Driver/intel	
3	LIBREOFFICE-45271	2012-01-26	Writer	
4	POLICYKIT-83093	2014-08-26	daemon	
5	PKG-CONFIG-32379	2010-12-14	src	
6	MESA-86357	2014-11-16	Drivers/Gallium/radeonsi	
7	POPPLER-75241	2014-02-20	utils	
8	XORG-2701	2005-03-11	Driver/intel	
9	XORG-15083	2008-03-16	App/other	
10	POPPLER-6199	2006-03-10	glib frontend	
11	XORG-21214	2009-04-15	Input/synaptics	
12	MESA-84807	2014-10-08	Mesa core	
13	LIBREOFFICE-33358	2011-01-21	Spreadsheet	
14	XORG-23838	2009-09-10	Driver/geode	

	product_name	short_description	\
0	FREEDESKTOP.ORG	New account request for mesa/r300/r600 develop…	
1	XORG	XChangeProperty prototype missing from manual …	
2	XORG	intel-virtual-output: fatal IO error 9 when ex…	
3	LIBREOFFICE	: The way Writer counting Chinese words is not…	

```
4         POLICYKIT       [patch] pkexec parameter parsing memory leak
5        PKG-CONFIG                                    c99ism in v0.25
6              MESA  [RadeonSI] GPU lockup with mesa 10.3.3 / kerne…
7           POPPLER  pdftops 0.25.1 -eps -level1sep creates an inva…
8              XORG  Switch to console or other session changes res…
9              XORG          xorg/app/xfontsel - Compile warning fixes
10          POPPLER                        mem leak in font properties
11             XORG  The synaptics edge scrolling defaults are too …
12             MESA  Build issue starting between bf4aecfb2acc8d0dc…
13      LIBREOFFICE      [FORMATTING] Wrong Russian format of the date
14             XORG        xf86-video-geode: git problem when updating

                                    long_description   assignee_name  \
0   Created attachment 29953\nssh-pub\n\nWould lik…   sitewranglers
1   The manual page for XChangeProperty lists seve…       xorg-team
2   Created attachment 100992\ni-v-o debug output\…           chris
3   Problem description: \n\nThe way of counting C…         caolanm
4   Google's Project Zero has developed an exploit…         zeuthen
5   Version 0.25 of pkg-config does not build on N…          tfheen
6   Created attachment 109569\njournalctl crash in…       dri-devel
7   pdftops -eps -level1sep creates an invalid EPS…    poppler-bugs
8   Using the current CVS sources for the i810 dri…           alanh
9   Created attachment 15214\n0001-Compile-warning…       xorg-team
10  font_info struct is not freed in poppler_font_…    poppler-bugs
11  For some touchpads (notably alps) the default …  peter.hutterer
12  Created attachment 107564\nexample patch\n\nHa…  emil.l.velikov
13  When I format cell as date by template 'D MMMM…           erack
14  Doing a git pull on a newly cloned repo always…       xorg-team

      reporter_name resolution_category  resolution_code  status_category  \
0         amaasikas               fixed                1                1
1    matthieu.herrb               fixed                1                1
2        main.haarp               fixed                1                1
3          blackjay               fixed                1                0
4             hanno               fixed                1                1
5             hauke               fixed                1                1
6          madcatx               fixed                1                1
7      williambader               fixed                1                1
8          sndirsch               fixed                1                0
9              pcpa               fixed                1                1
10         carlosgc               fixed                1                1
11  peter.hutterer               fixed                1                1
12       adf.lists               fixed                1                1
13  mistresssilvara               fixed                1                1
14         memsize               fixed                1                1

    status_code update_date  quantity_of_votes  quantity_of_comments  \
```

```
0              4  2009-09-30                  0                     4
1              4  2004-12-14                  0                     5
2              4  2014-06-13                  0                    10
3              6  2012-04-05                  0                     4
4              4  2014-08-27                  0                     6
5              4  2011-04-13                  0                     2
6              4  2015-08-02                  0                     7
7              4  2014-03-15                  0                     6
8              6  2005-03-14                  0                    48
9              4  2008-11-27                  0                     2
10             4  2006-03-20                  0                     3
11             4  2009-05-04                  0                     6
12             4  2014-10-14                  0                     8
13             4  2011-12-09                  0                     4
14             4  2010-04-14                  0                     6

   resolution_date  bug_fix_time  severity_category  severity_code
0       2009-09-30             0                  0              2
1       2004-12-14             2                  0              2
2       2014-06-13             0                  0              2
3       2012-04-05            70                  4              4
4       2014-08-27             1                  0              2
5       2011-04-13           120                  0              2
6       2015-08-02           259                  0              2
7       2014-03-15            23                  0              2
8       2005-03-15             4                  0              2
9       2008-11-27           256                  0              2
10      2006-03-21            11                  0              2
11      2009-05-04            19                  0              2
12      2014-10-14             6                  0              2
13      2011-12-09           322                  0              2
14      2010-04-14           216                  0              2
```

[ ]:
```
fbrd['component_name'].unique
```

[ ]:
```
<bound method Series.unique of 0        New Accounts
1       Documentation
2        Driver/intel
3              Writer
4              daemon
             …
7679           weston
7680           Lib/Xt
7681     Driver/intel
7682             core
7683          General
Name: component_name, Length: 7684, dtype: object>
```

[ ]:
```
fbrd['product_name'].unique
```

[ ]:
```
<bound method Series.unique of 0         FREEDESKTOP.ORG
1                   XORG
2                   XORG
3             LIBREOFFICE
4               POLICYKIT
               …
7679              WAYLAND
7680                 XORG
7681                 XORG
7682            PULSEAUDIO
7683    MEDIA-PLAYER-INFO
Name: product_name, Length: 7684, dtype: object>
```

1.3 As other elements are not having a categorical

1.3.1 Now we will be removing null values

[ ]:
```
fbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description       48
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
# percentage of missing values
fbrd.isnull().sum() / fbrd.shape[0] * 100
```

[ ]:
```
bug_id                  0.000000
creation_date           0.000000
```

```
component_name          0.000000
product_name            0.000000
short_description       0.000000
long_description        0.624675
assignee_name           0.000000
reporter_name           0.000000
resolution_category     0.000000
resolution_code         0.000000
status_category         0.000000
status_code             0.000000
update_date             0.000000
quantity_of_votes       0.000000
quantity_of_comments    0.000000
resolution_date         0.000000
bug_fix_time            0.000000
severity_category       0.000000
severity_code           0.000000
dtype: float64
```

[ ]:
```
fbrd.dropna(inplace=True)
```

[ ]:
```
fbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description        0
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
fbrd.head()
```

[ ]:

```
                  bug_id creation_date component_name     product_name  \
0  FREEDESKTOP.ORG-24230    2009-09-30   New Accounts  FREEDESKTOP.ORG
1              XORG-2056    2004-12-12  Documentation             XORG
2             XORG-79994    2014-06-13   Driver/intel             XORG
3     LIBREOFFICE-45271    2012-01-26         Writer      LIBREOFFICE
4       POLICYKIT-83093    2014-08-26         daemon        POLICYKIT

                                  short_description  \
0  New account request for mesa/r300/r600 develop…
1  XChangeProperty prototype missing from manual …
2  intel-virtual-output: fatal IO error 9 when ex…
3  : The way Writer counting Chinese words is not…
4       [patch] pkexec parameter parsing memory leak

                                   long_description  assignee_name  \
0  Created attachment 29953\nssh-pub\n\nWould lik…  sitewranglers
1  The manual page for XChangeProperty lists seve…      xorg-team
2  Created attachment 100992\ni-v-o debug output\…          chris
3  Problem description: \n\nThe way of counting C…        caolanm
4  Google's Project Zero has developed an exploit…        zeuthen

    reporter_name resolution_category  resolution_code  status_category  \
0       amaasikas               fixed               1                1
1  matthieu.herrb               fixed               1                1
2      main.haarp               fixed               1                1
3        blackjay               fixed               1                0
4           hanno               fixed               1                1

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            4  2009-09-30                  0                     4
1            4  2004-12-14                  0                     5
2            4  2014-06-13                  0                    10
3            6  2012-04-05                  0                     4
4            4  2014-08-27                  0                     6

  resolution_date  bug_fix_time  severity_category  severity_code
0      2009-09-30             0                  0              2
1      2004-12-14             2                  0              2
2      2014-06-13             0                  0              2
3      2012-04-05            70                  4              4
4      2014-08-27             1                  0              2
```

[ ]:

```
from datetime import time,date,datetime

fbrd['creation_date'] = pd.to_datetime(fbrd['creation_date'])
fbrd['update_date'] = pd.to_datetime(fbrd['update_date'])
fbrd['resolution_date'] = pd.to_datetime(fbrd['resolution_date'])
```

[ ]:
```
ebrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9627 entries, 0 to 9775
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   time_to_dep[s]        9627 non-null   float64
 1   status_category       9627 non-null   int64
 2   status_code           9627 non-null   int64
 3   quantity_of_votes     9627 non-null   int64
 4   quantity_of_comments  9627 non-null   int64
 5   bug_fix_time          9627 non-null   int64
 6   severity_category     9627 non-null   int64
 7   severity_code         9627 non-null   int64
dtypes: float64(1), int64(7)
memory usage: 676.9 KB
```

[ ]:
```
fbrd.insert(0,"time_to_dep[s]",␣
↪((fbrd['resolution_date']-fbrd['creation_date']).astype('timedelta64[s]')),␣
↪True)
```

[ ]:
```
fbrd.head()
```

[ ]:

	time_to_dep[s]	bug_id	creation_date	component_name	product_name	short_description	long_description	assignee_name	reporter_name	resolution_category	resolution_code	status_category
0	0.0	FREEDESKTOP.ORG-24230	2009-09-30	New Accounts	FREEDESKTOP.ORG	New account request for mesa/r300/r600 develop…	Created attachment 29953\nssh-pub\n\nWould lik…	sitewranglers	amaasikas	fixed	1	1
1	172800.0	XORG-2056	2004-12-12	Documentation	XORG	XChangeProperty prototype missing from manual …	The manual page for XChangeProperty lists seve…	xorg-team				
2	0.0	XORG-79994	2014-06-13	Driver/intel	XORG	intel-virtual-output: fatal IO error 9 when ex…	Created attachment 100992\ni-v-o debug output\…	chris				
3	6048000.0	LIBREOFFICE-45271	2012-01-26	Writer	LIBREOFFICE	: The way Writer counting Chinese words is not…	Problem description: \n\nThe way of counting C…	caolanm				
4	86400.0	POLICYKIT-83093	2014-08-26	daemon	POLICYKIT	[patch] pkexec parameter parsing memory leak	Google's Project Zero has developed an exploit…	zeuthen				

```
1  matthieu.herrb              fixed                1                  1
2      main.haarp              fixed                1                  1
3        blackjay              fixed                1                  0
4           hanno              fixed                1                  1

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            4  2009-09-30                  0                     4
1            4  2004-12-14                  0                     5
2            4  2014-06-13                  0                    10
3            6  2012-04-05                  0                     4
4            4  2014-08-27                  0                     6

  resolution_date  bug_fix_time  severity_category  severity_code
0      2009-09-30             0                  0              2
1      2004-12-14             2                  0              2
2      2014-06-13             0                  0              2
3      2012-04-05            70                  4              4
4      2014-08-27             1                  0              2
```

[ ]:
```
# dropping extra columns
fbrd.
 ↪drop(['bug_id','resolution_code','quantity_of_votes','creation_date','component_name','prod
```

[ ]:
```
# now we are done with dropping values as well
fbrd['bug_fix_time'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew         4.936793
kurtosis    35.151139
Name: bug_fix_time, dtype: float64
```

[ ]:
```
corr = fbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.050669    -0.050669
status_category             0.050669         1.000000    -1.000000
status_code                -0.050669        -1.000000     1.000000
quantity_of_comments        0.131916        -0.095479     0.095479
bug_fix_time                1.000000         0.050669    -0.050669
severity_category           0.000149        -0.087373     0.087373
severity_code              -0.034816        -0.123314     0.123314

                      quantity_of_comments  bug_fix_time  severity_category  \
time_to_dep[s]                    0.131916      1.000000           0.000149
status_category                  -0.095479      0.050669          -0.087373
status_code                       0.095479     -0.050669           0.087373
quantity_of_comments              1.000000      0.131916           0.127924
```

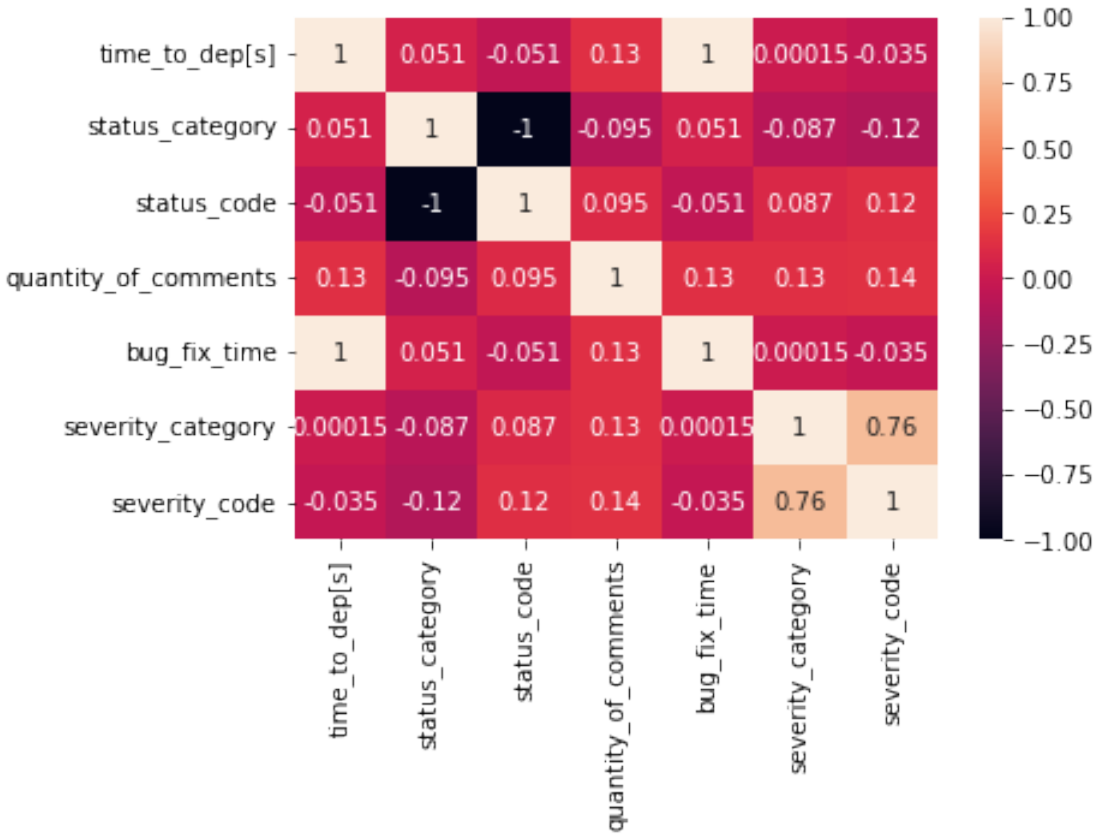
```
bug_fix_time                  0.131916      1.000000           0.000149
severity_category             0.127924      0.000149           1.000000
severity_code                 0.136716     -0.034816           0.757478

                      severity_code
time_to_dep[s]            -0.034816
status_category           -0.123314
status_code                0.123314
quantity_of_comments       0.136716
bug_fix_time              -0.034816
severity_category          0.757478
severity_code              1.000000
```

[ ]:
```
corr = fbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.050669    -0.050669
status_category             0.050669         1.000000    -1.000000
status_code                -0.050669        -1.000000     1.000000
quantity_of_comments        0.131916        -0.095479     0.095479
bug_fix_time                1.000000         0.050669    -0.050669
severity_category           0.000149        -0.087373     0.087373
severity_code              -0.034816        -0.123314     0.123314

                      quantity_of_comments  bug_fix_time  severity_category  \
time_to_dep[s]                    0.131916      1.000000           0.000149
status_category                  -0.095479      0.050669          -0.087373
status_code                       0.095479     -0.050669           0.087373
quantity_of_comments              1.000000      0.131916           0.127924
bug_fix_time                      0.131916      1.000000           0.000149
severity_category                 0.127924      0.000149           1.000000
severity_code                     0.136716     -0.034816           0.757478

                      severity_code
time_to_dep[s]            -0.034816
status_category           -0.123314
status_code                0.123314
quantity_of_comments       0.136716
bug_fix_time              -0.034816
severity_category          0.757478
severity_code              1.000000
```

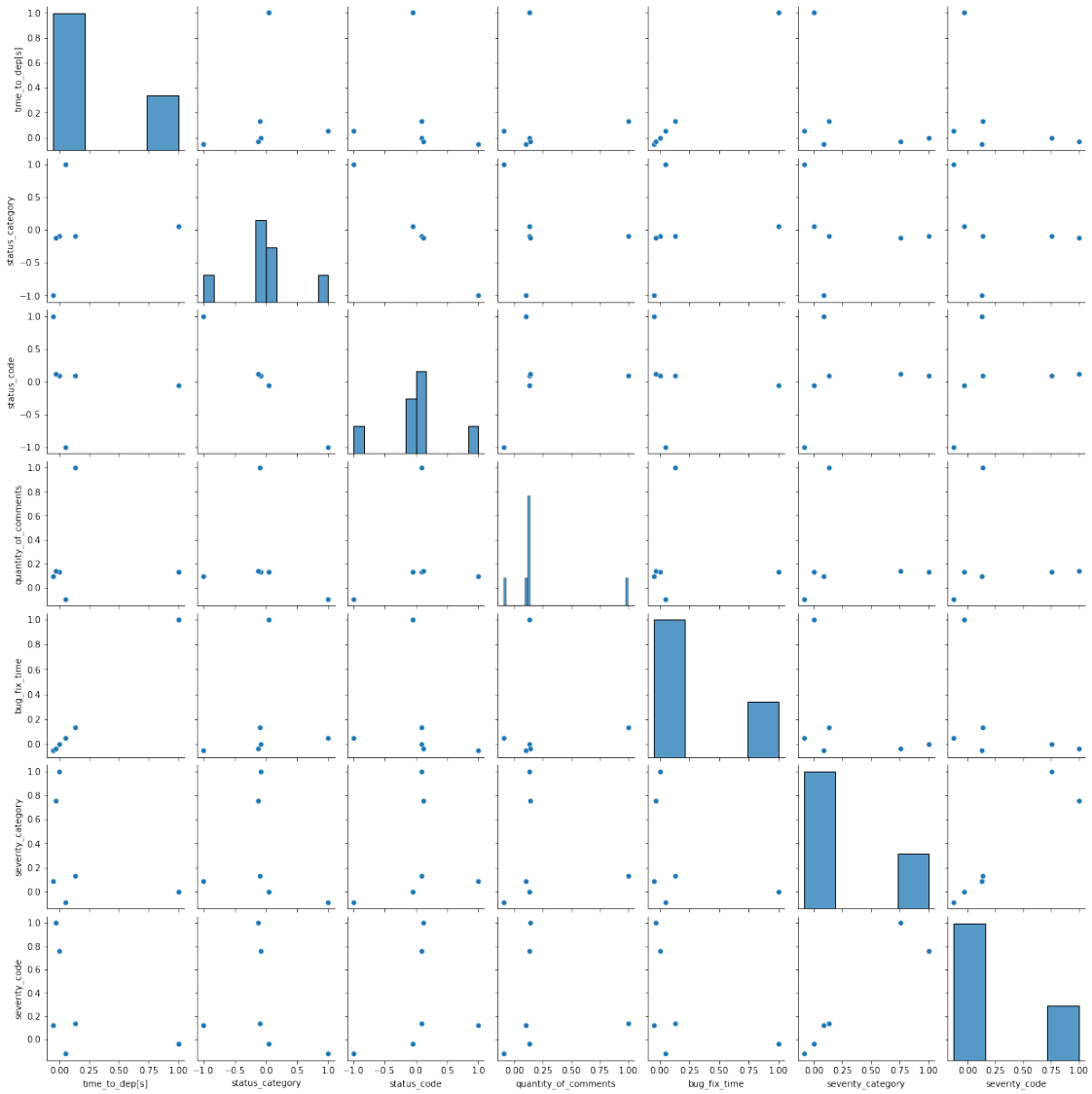
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]:
```
<AxesSubplot:>
```



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]:
```
<seaborn.axisgrid.PairGrid at 0x266a8589090>
```



[ ]:
```
fbrd.head()
```

[ ]:

	time_to_dep[s]	status_category	status_code	quantity_of_comments	\
0	0.0	1	4	4	
1	172800.0	1	4	5	
2	0.0	1	4	10	
3	6048000.0	0	6	4	
4	86400.0	1	4	6	

	bug_fix_time	severity_category	severity_code
0	0	0	2
1	2	0	2
2	0	0	2

```
3            70                 4              4
4             1                 0              2
```

[ ]:
```
fbrd.tail()
```

[ ]:
```
      time_to_dep[s]  status_category  status_code  quantity_of_comments  \
7679      37929600.0                1            4                     3
7680         86400.0                1            4                     8
7681      13046400.0                1            4                    24
7682         86400.0                1            4                     2
7683             0.0                1            4                     2

      bug_fix_time  severity_category  severity_code
7679           439                  4              4
7680             1                  4              4
7681           151                  4              4
7682             1                  0              2
7683             0                  0              2
```

[ ]:
```
X = fbrd.iloc[:,:-1].values #rows and then columns in brackets
Y = fbrd.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[0.00000e+00, 1.00000e+00, 4.00000e+00, 4.00000e+00, 0.00000e+00,
        0.00000e+00],
       [1.72800e+05, 1.00000e+00, 4.00000e+00, 5.00000e+00, 2.00000e+00,
        0.00000e+00],
       [0.00000e+00, 1.00000e+00, 4.00000e+00, 1.00000e+01, 0.00000e+00,
        0.00000e+00],
       …,
       [1.30464e+07, 1.00000e+00, 4.00000e+00, 2.40000e+01, 1.51000e+02,
        4.00000e+00],
       [8.64000e+04, 1.00000e+00, 4.00000e+00, 2.00000e+00, 1.00000e+00,
        0.00000e+00],
       [0.00000e+00, 1.00000e+00, 4.00000e+00, 2.00000e+00, 0.00000e+00,
        0.00000e+00]])
```

[ ]:
```
Y
```

[ ]:
```
array([2, 2, 2, …, 4, 2, 2], dtype=int64)
```

1.3.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=1234)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
```

```
LinearRegression()
R2 score is:  0.5853911475716607
Mean Absolute error is:  0.26820646356298156
Mean Squared Error:  0.5214388436143407
---------------------------------------------------
LogisticRegression()
R2 score is:  -0.20985909497254296
Mean Absolute error is:  0.5281413612565445
Mean Squared Error:  1.5215968586387434
---------------------------------------------------
GaussianNB()
R2 score is:  -0.6053399174151806
Mean Absolute error is:  0.6982984293193717
```

```
Mean Squared Error:  2.018979057591623
--------------------------------------------------
RandomForestClassifier()
R2 score is:  0.9620130262653782
Mean Absolute error is:  0.01112565445026178
Mean Squared Error:  0.047774869109947646
--------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
--------------------------------------------------
SVR()
R2 score is:  -0.13615953475606557
Mean Absolute error is:  0.5894829423470678
Mean Squared Error:  1.4289075365726942
--------------------------------------------------
```

1.3.3 EDA On gcc_bug_report_data

[ ]:
```
gcbrd.head()
```

[ ]:
```
      bug_id creation_date component_name product_name  \
0  GCC-49282    2011-06-04     middle-end          GCC
1  GCC-36574    2008-06-19     middle-end          GCC
2  GCC-77269    2016-08-16     middle-end          GCC
3  GCC-78479    2016-11-22        fortran          GCC
4    GCC-632    2000-10-12            c++          GCC

                                  short_description  \
0  malloc corruption in large lto1-wpa run during…
1  [4.4 Regression] build broken with cgraph changes
2  __builtin_isinf_sign does not work for __float128
3       ICE in gfc_apply_init at fortran/expr.c:4135
4  Internal compiler error in `layout_decl' at st…

                                  long_description assignee_name  \
0  A large lto1-wpa run with 20110603 results now…    unassigned
1  With r136888 cris-elf built (and had just 4 re…    unassigned
2  The __builtin_isinf_sign folding does\n\n     …         jsm28
3  With valid code down to at least 4.8 :\n\n\n$ …         kargl
4  g++  -I/home/leila/Sources/Include -O2 -Wall -…    unassigned

               reporter_name resolution_category  resolution_code  \
0                  andi-gcc               fixed                1
1                         hp              fixed                1
2                     jsm28               fixed                1
```

```
3  gerhard.steinmetz.fortran              fixed                1
4                        lkh              fixed                1

  status_category  status_code update_date  quantity_of_votes  \
0        resolved            4  2011-10-07                  0
1        resolved            4  2008-07-18                  0
2        resolved            4  2016-08-22                  0
3        resolved            4  2016-11-22                  0
4        resolved            4  2003-07-25                  0

   quantity_of_comments resolution_date  bug_fix_time severity_category  \
0                     9      2011-10-07           125            normal
1                     5      2008-07-18            29            normal
2                     5      2016-08-22             6            normal
3                     6      2016-11-22             0            normal
4                     5      2001-02-04           115            normal

   severity_code
0              2
1              2
2              2
3              2
4              2
```

[ ]:
```
gcbrd.tail()
```

[ ]:
```
         bug_id creation_date component_name product_name  \
9995  GCC-58027    2013-07-30        fortran          GCC
9996   GCC-5094    2001-12-12            c++          GCC
9997  GCC-57073    2013-04-25     middle-end          GCC
9998  GCC-84112    2018-01-29         target          GCC
9999  GCC-69432    2016-01-22          other          GCC

                                  short_description  \
9995  Arithmetic overflow converting … in PARAMETE…
9996          partial specialisation cannot be friend??
9997  __builtin_powif (-1.0 k) should be optimized t…
9998   [8 Regression] powerpc64le ICE in LRA on openjdk
9999  [4.9 Regression] ICE in connect_traces at dwar…

                                  long_description assignee_name  \
9995  Compiling the following code (extracted from h…         kargl
9996  The attached file tries several configuration …       lerdsuwa
9997  Motivated by PR57071.\n\nIn numerical code it …    unassigned
9998  The following testcase ICEs with -mcpu=power8 …    unassigned
9999  Created attachment 37434\nC-reduced testcase\n…         jakub
```

```
     reporter_name resolution_category  resolution_code status_category  \
9995       dominiq               fixed               1        resolved
9996         benko               fixed               1        resolved
9997        burnus               fixed               1        resolved
9998         jakub               fixed               1        resolved
9999       jamborm               fixed               1        resolved

      status_code update_date  quantity_of_votes  quantity_of_comments  \
9995            4  2015-11-17                  0                     6
9996            4  2005-06-18                  0                     7
9997            4  2013-06-01                  0                    17
9998            4  2018-01-31                  0                     4
9999            4  2016-02-11                  0                     9

     resolution_date  bug_fix_time severity_category  severity_code
9995      2015-11-17           840            normal              2
9996      2002-12-19           372            normal              2
9997      2013-06-01            37            normal              2
9998      2018-01-31             2            normal              2
9999      2016-02-11            20            normal              2
```

[ ]:
```
gcbrd.describe()
```

[ ]:
```
       resolution_code   status_code  quantity_of_votes  quantity_of_comments  \
count          10000.0  10000.000000            10000.0          10000.000000
mean               1.0      4.019400                0.0              8.145000
std                0.0      0.196029                0.0              6.553859
min                1.0      4.000000                0.0              1.000000
25%                1.0      4.000000                0.0              4.000000
50%                1.0      4.000000                0.0              6.000000
75%                1.0      4.000000                0.0             10.000000
max                1.0      6.000000                0.0            124.000000

       bug_fix_time  severity_code
count    10000.0000    10000.00000
mean       225.8268        2.22120
std        516.6366        0.79758
min       -527.0000        1.00000
25%          3.0000        2.00000
50%         27.0000        2.00000
75%        188.0000        2.00000
max       5351.0000        6.00000
```

[ ]:
```
gcbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
```

```
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                10000 non-null  object
 1   creation_date         10000 non-null  object
 2   component_name        10000 non-null  object
 3   product_name          10000 non-null  object
 4   short_description     9994 non-null   object
 5   long_description      9986 non-null   object
 6   assignee_name         10000 non-null  object
 7   reporter_name         10000 non-null  object
 8   resolution_category   10000 non-null  object
 9   resolution_code       10000 non-null  int64
 10  status_category       10000 non-null  object
 11  status_code           10000 non-null  int64
 12  update_date           10000 non-null  object
 13  quantity_of_votes     10000 non-null  int64
 14  quantity_of_comments  10000 non-null  int64
 15  resolution_date       10000 non-null  object
 16  bug_fix_time          10000 non-null  int64
 17  severity_category     10000 non-null  object
 18  severity_code         10000 non-null  int64
dtypes: int64(6), object(13)
memory usage: 1.4+ MB
```

[ ]:
```
gcbrd['status_category'].unique()
```

[ ]:
```
array(['resolved', 'closed'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status_category we can encode it as it is two categories only
# We are encoding it
gcbrd['status_category'] = gcbrd['status_category'].replace("resolved", 1)
gcbrd['status_category'] = gcbrd['status_category'].replace("closed", 0)
```

[ ]:
```
gcbrd['severity_category'].unique()
```

[ ]:
```
array(['normal', 'critical', 'minor', 'blocker', 'major', 'trivial'],
      dtype=object)
```

[ ]:
```
# As we have another column severity_category and we can perform encoding on it␣
 ↪as well
gcbrd['severity_category'] = gcbrd['severity_category'].replace("normal", 0)
gcbrd['severity_category'] = gcbrd['severity_category'].replace("blocker", 1)
gcbrd['severity_category'] = gcbrd['severity_category'].replace("trivial", 2)
gcbrd['severity_category'] = gcbrd['severity_category'].replace("minor", 3)
gcbrd['severity_category'] = gcbrd['severity_category'].replace("major", 4)
```

```
gcbrd['severity_category'] = gcbrd['severity_category'].replace("critical", 5)
```

[ ]:
```
gcbrd.head(15)
```

[ ]:
```
             bug_id creation_date     component_name product_name  \
0         GCC-49282    2011-06-04        middle-end          GCC
1         GCC-36574    2008-06-19        middle-end          GCC
2         GCC-77269    2016-08-16        middle-end          GCC
3         GCC-78479    2016-11-22           fortran          GCC
4           GCC-632    2000-10-12               c++          GCC
5          GCC-4071    2001-08-21         libstdc++          GCC
6         GCC-67037    2015-07-27  rtl-optimization          GCC
7         GCC-83112    2017-11-22            libgcc          GCC
8          GCC-7111    2002-06-24         libstdc++          GCC
9         GCC-18737    2004-11-30           fortran          GCC
10        GCC-58603    2013-10-03            target          GCC
11        GCC-38480    2008-12-10  tree-optimization          GCC
12        GCC-58545    2013-09-26   rtl-optimization          GCC
13  CLASSPATH-16983    2004-08-11             swing    CLASSPATH
14        GCC-48795    2011-04-27  tree-optimization          GCC

                                    short_description  \
0   malloc corruption in large lto1-wpa run during…
1   [4.4 Regression] build broken with cgraph changes
2   __builtin_isinf_sign does not work for __float128
3        ICE in gfc_apply_init at fortran/expr.c:4135
4   Internal compiler error in `layout_decl' at st…
5   make install installs …/include/g++-v3/Makefile
6   [4.9 Regression] Wrong code at -O1 and above o…
7   Silence warnings from PowerPC libgcc float128-…
8               cout of null pointer causes core dump
9              ICE on invalid use of external keyword
10  [4.9 Regression] hash-table.h:962: error: anac…
11  bogus warning with -O3 -Wall: array subscript …
12  [4.8 Regression] error: unable to find a regis…
13                           ButtonGroups don't work
14                     -Warray-bounds false positive

                                    long_description   assignee_name  \
0   A large lto1-wpa run with 20110603 results now…      unassigned
1   With r136888 cris-elf built (and had just 4 re…      unassigned
2   The __builtin_isinf_sign folding does\n\n     …           jsm28
3   With valid code down to at least 4.8 :\n\n\n$ …           kargl
4   g++  -I/home/leila/Sources/Include -O2 -Wall -…      unassigned
5   [I couldn't work out which category/class prob…            bkoz
6   Created attachment 36076\ntestcase\n\nThe redu…  bernd.edlinger
7   The ifunc handlers in libgcc to switch between…        meissner
```

```
8   Trying to print a null character pointer cause…            bkoz
9   $ cat a.f90 \nprogram test\n  implicit none\n …      unassigned
10  In stage1 with gcc 4.4.7 build fails:\n\ng++ -…      unassigned
11  # gcc -v\nUsing built-in specs.\nTarget: i486-…      unassigned
12  == C source ==\n\ntypedef unsigned char uint8_…      unassigned
13                                  Working on it.             kho
14  $ gcc -v\nUtilisation des specs internes.\nCOL…         rguenth

              reporter_name resolution_category  resolution_code  \
0                  andi-gcc               fixed                1
1                       hp                fixed                1
2                    jsm28                fixed                1
3   gerhard.steinmetz.fortran               fixed                1
4                       lkh                 fixed                1
5                       pgr                 fixed                1
6                   notasas                 fixed                1
7                   meissner                fixed                1
8                     rdhunt                fixed                1
9                   fxcoudert               fixed                1
10                  danglin                 fixed                1
11                  helmert                 fixed                1
12                      gjl                 fixed                1
13                      kho                 fixed                1
14               niko.lecam                 fixed                1

    status_category  status_code update_date  quantity_of_votes  \
0                 1            4  2011-10-07                  0
1                 1            4  2008-07-18                  0
2                 1            4  2016-08-22                  0
3                 1            4  2016-11-22                  0
4                 1            4  2003-07-25                  0
5                 1            4  2003-07-25                  0
6                 1            4  2016-08-03                  0
7                 1            4  2017-12-01                  0
8                 1            4  2004-01-02                  0
9                 1            4  2005-11-01                  0
10                1            4  2013-10-19                  0
11                1            4  2017-10-17                  0
12                1            4  2015-06-23                  0
13                1            4  2005-10-16                  0
14                1            4  2016-02-25                  0

    quantity_of_comments resolution_date  bug_fix_time  severity_category  \
0                      9      2011-10-07           125                  0
1                      5      2008-07-18            29                  0
2                      5      2016-08-22             6                  0
3                      6      2016-11-22             0                  0
```

```
4                     5      2001-02-04          115                   0
5                     7      2001-10-10           50                   0
6                    14      2016-08-03          373                   0
7                     4      2017-11-30            8                   0
8                     6      2002-07-04           10                   0
9                     9      2005-11-01          336                   0
10                    9      2013-10-19           16                   0
11                    4      2017-10-17         3233                   0
12                   11      2015-06-23          635                   0
13                    3      2004-08-26           15                   0
14                    9      2016-02-25         1765                   0

    severity_code
0               2
1               2
2               2
3               2
4               2
5               2
6               2
7               2
8               2
9               2
10              2
11              2
12              2
13              2
14              2
```

[ ]:
```
gcbrd['component_name'].unique
```

[ ]:
```
<bound method Series.unique of 0       middle-end
1       middle-end
2       middle-end
3          fortran
4              c++
           …
9995       fortran
9996           c++
9997    middle-end
9998        target
9999         other
Name: component_name, Length: 10000, dtype: object>
```

[ ]:
```
gcbrd['product_name'].unique
```

[ ]:
```
<bound method Series.unique of 0       GCC
1       GCC
2       GCC
3       GCC
4       GCC
       …
9995    GCC
9996    GCC
9997    GCC
9998    GCC
9999    GCC
Name: product_name, Length: 10000, dtype: object>
```

1.4 As other elements are not having a categorical

1.4.1 Now we will be removing null values

[ ]:
```
gcbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       6
long_description       14
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
# percentage of missing values
gcbrd.isnull().sum() / gcbrd.shape[0] * 100
```

[ ]:
```
bug_id                  0.00
creation_date           0.00
component_name          0.00
product_name            0.00
```

```
short_description       0.06
long_description        0.14
assignee_name           0.00
reporter_name           0.00
resolution_category     0.00
resolution_code         0.00
status_category         0.00
status_code             0.00
update_date             0.00
quantity_of_votes       0.00
quantity_of_comments    0.00
resolution_date         0.00
bug_fix_time            0.00
severity_category       0.00
severity_code           0.00
dtype: float64
```

[ ]:
```
gcbrd.dropna(inplace=True)
```

[ ]:
```
gcbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description        0
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
gcbrd.head()
```

[ ]:
```
       bug_id creation_date component_name product_name  \
0  GCC-49282    2011-06-04     middle-end          GCC
1  GCC-36574    2008-06-19     middle-end          GCC
```

```
2  GCC-77269    2016-08-16     middle-end          GCC
3  GCC-78479    2016-11-22        fortran          GCC
4    GCC-632    2000-10-12            c++          GCC

                                 short_description  \
0  malloc corruption in large lto1-wpa run during…
1  [4.4 Regression] build broken with cgraph changes
2  __builtin_isinf_sign does not work for __float128
3       ICE in gfc_apply_init at fortran/expr.c:4135
4  Internal compiler error in `layout_decl' at st…

                                  long_description assignee_name  \
0  A large lto1-wpa run with 20110603 results now…    unassigned
1  With r136888 cris-elf built (and had just 4 re…    unassigned
2  The __builtin_isinf_sign folding does\n\n     …         jsm28
3  With valid code down to at least 4.8 :\n\n\n$ …         kargl
4  g++  -I/home/leila/Sources/Include -O2 -Wall -…    unassigned

              reporter_name resolution_category  resolution_code  \
0                  andi-gcc              fixed                1
1                        hp              fixed                1
2                     jsm28              fixed                1
3  gerhard.steinmetz.fortran              fixed                1
4                       lkh              fixed                1

   status_category  status_code update_date  quantity_of_votes  \
0                1            4  2011-10-07                  0
1                1            4  2008-07-18                  0
2                1            4  2016-08-22                  0
3                1            4  2016-11-22                  0
4                1            4  2003-07-25                  0

   quantity_of_comments resolution_date  bug_fix_time  severity_category  \
0                     9      2011-10-07           125                  0
1                     5      2008-07-18            29                  0
2                     5      2016-08-22             6                  0
3                     6      2016-11-22             0                  0
4                     5      2001-02-04           115                  0

   severity_code
0              2
1              2
2              2
3              2
4              2
```

[ ]:
```
from datetime import time,date,datetime

gcbrd['creation_date'] = pd.to_datetime(gcbrd['creation_date'])
gcbrd['update_date'] = pd.to_datetime(gcbrd['update_date'])
gcbrd['resolution_date'] = pd.to_datetime(gcbrd['resolution_date'])
```

[ ]:
```
gcbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9980 entries, 0 to 9999
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                9980 non-null   object
 1   creation_date         9980 non-null   datetime64[ns]
 2   component_name        9980 non-null   object
 3   product_name          9980 non-null   object
 4   short_description     9980 non-null   object
 5   long_description      9980 non-null   object
 6   assignee_name         9980 non-null   object
 7   reporter_name         9980 non-null   object
 8   resolution_category   9980 non-null   object
 9   resolution_code       9980 non-null   int64
 10  status_category       9980 non-null   int64
 11  status_code           9980 non-null   int64
 12  update_date           9980 non-null   datetime64[ns]
 13  quantity_of_votes     9980 non-null   int64
 14  quantity_of_comments  9980 non-null   int64
 15  resolution_date       9980 non-null   datetime64[ns]
 16  bug_fix_time          9980 non-null   int64
 17  severity_category     9980 non-null   int64
 18  severity_code         9980 non-null   int64
dtypes: datetime64[ns](3), int64(8), object(8)
memory usage: 1.5+ MB
```

[ ]:
```
gcbrd.insert(0,"time_to_dep[s]",␣
 ↪((gcbrd['resolution_date']-gcbrd['creation_date']).
 ↪astype('timedelta64[s]')), True)
```

[ ]:
```
gcbrd.head()
```

[ ]:

	time_to_dep[s]	bug_id	creation_date	component_name	product_name	\
0	10800000.0	GCC-49282	2011-06-04	middle-end	GCC	
1	2505600.0	GCC-36574	2008-06-19	middle-end	GCC	
2	518400.0	GCC-77269	2016-08-16	middle-end	GCC	
3	0.0	GCC-78479	2016-11-22	fortran	GCC	
4	9936000.0	GCC-632	2000-10-12	c++	GCC	

```
                                  short_description  \
0  malloc corruption in large lto1-wpa run during…
1  [4.4 Regression] build broken with cgraph changes
2  __builtin_isinf_sign does not work for __float128
3       ICE in gfc_apply_init at fortran/expr.c:4135
4  Internal compiler error in `layout_decl' at st…

                                   long_description assignee_name  \
0  A large lto1-wpa run with 20110603 results now…    unassigned
1  With r136888 cris-elf built (and had just 4 re…    unassigned
2  The __builtin_isinf_sign folding does\n\n     …          jsm28
3  With valid code down to at least 4.8 :\n\n\n$ …          kargl
4  g++  -I/home/leila/Sources/Include -O2 -Wall -…    unassigned

              reporter_name resolution_category  resolution_code  \
0                 andi-gcc               fixed                1
1                       hp               fixed                1
2                   jsm28               fixed                1
3  gerhard.steinmetz.fortran               fixed                1
4                       lkh               fixed                1

   status_category  status_code update_date  quantity_of_votes  \
0                1            4  2011-10-07                  0
1                1            4  2008-07-18                  0
2                1            4  2016-08-22                  0
3                1            4  2016-11-22                  0
4                1            4  2003-07-25                  0

   quantity_of_comments resolution_date  bug_fix_time  severity_category  \
0                     9      2011-10-07           125                  0
1                     5      2008-07-18            29                  0
2                     5      2016-08-22             6                  0
3                     6      2016-11-22             0                  0
4                     5      2001-02-04           115                  0

   severity_code
0              2
1              2
2              2
3              2
4              2
```

[ ]:
```
# dropping extra columns
gcbrd.
  ↪drop(['bug_id','resolution_code','quantity_of_votes','creation_date','component_name','prod
```

[ ]:
```
# now we are done with dropping values as well
gcbrd['bug_fix_time'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew         4.100914
kurtosis    21.124014
Name: bug_fix_time, dtype: float64
```

[ ]:
```
corr = gcbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.010646    -0.010646
status_category             0.010646         1.000000    -1.000000
status_code                -0.010646        -1.000000     1.000000
quantity_of_comments        0.138974         0.045323    -0.045323
bug_fix_time                1.000000         0.010646    -0.010646
severity_category          -0.028085         0.005807    -0.005807
severity_code              -0.076741         0.001341    -0.001341

                      quantity_of_comments  bug_fix_time  severity_category  \
time_to_dep[s]                    0.138974      1.000000          -0.028085
status_category                   0.045323      0.010646           0.005807
status_code                      -0.045323     -0.010646          -0.005807
quantity_of_comments              1.000000      0.138974           0.026818
bug_fix_time                      0.138974      1.000000          -0.028085
severity_category                 0.026818     -0.028085           1.000000
severity_code                     0.057272     -0.076741           0.831194

                      severity_code
time_to_dep[s]            -0.076741
status_category            0.001341
status_code               -0.001341
quantity_of_comments       0.057272
bug_fix_time              -0.076741
severity_category          0.831194
severity_code              1.000000
```

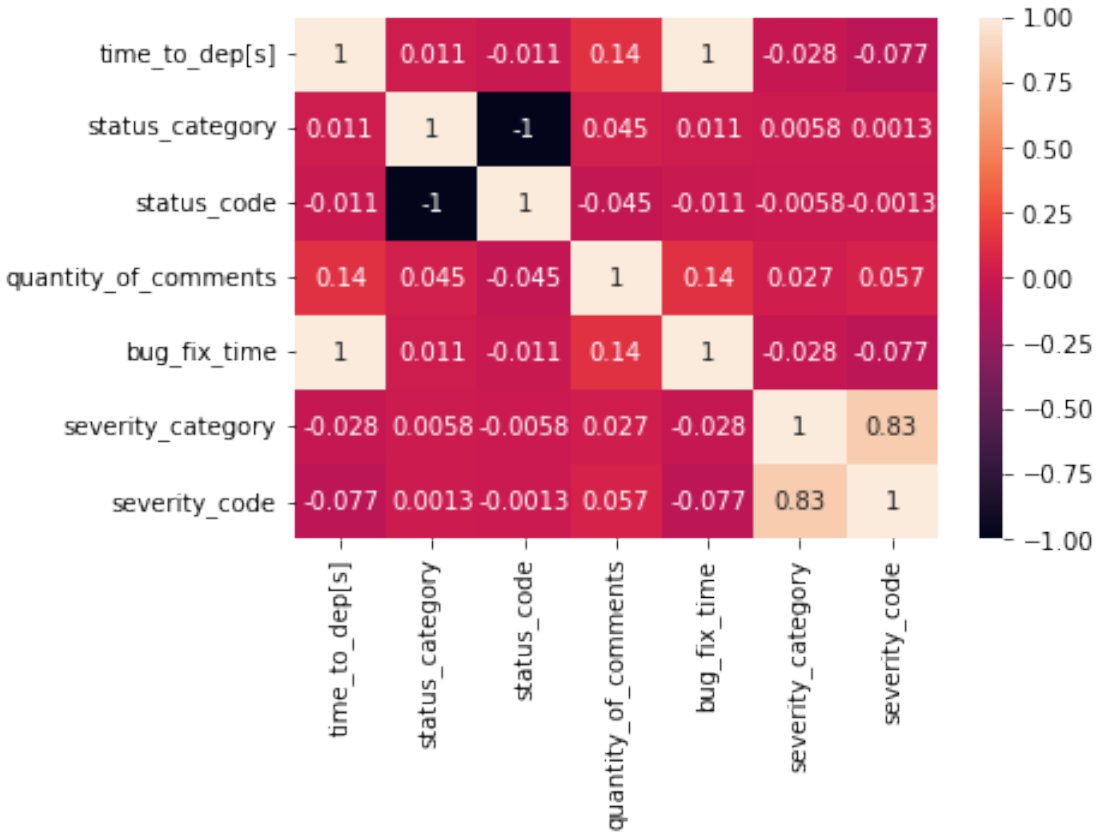
[ ]:
```
corr = gcbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.010646    -0.010646
status_category             0.010646         1.000000    -1.000000
status_code                -0.010646        -1.000000     1.000000
quantity_of_comments        0.138974         0.045323    -0.045323
bug_fix_time                1.000000         0.010646    -0.010646
severity_category          -0.028085         0.005807    -0.005807
```

```
severity_code               -0.076741         0.001341    -0.001341

                      quantity_of_comments  bug_fix_time  severity_category  \
time_to_dep[s]                    0.138974      1.000000          -0.028085
status_category                   0.045323      0.010646           0.005807
status_code                      -0.045323     -0.010646          -0.005807
quantity_of_comments              1.000000      0.138974           0.026818
bug_fix_time                      0.138974      1.000000          -0.028085
severity_category                 0.026818     -0.028085           1.000000
severity_code                     0.057272     -0.076741           0.831194

                      severity_code
time_to_dep[s]            -0.076741
status_category            0.001341
status_code               -0.001341
quantity_of_comments       0.057272
bug_fix_time              -0.076741
severity_category          0.831194
severity_code              1.000000
```

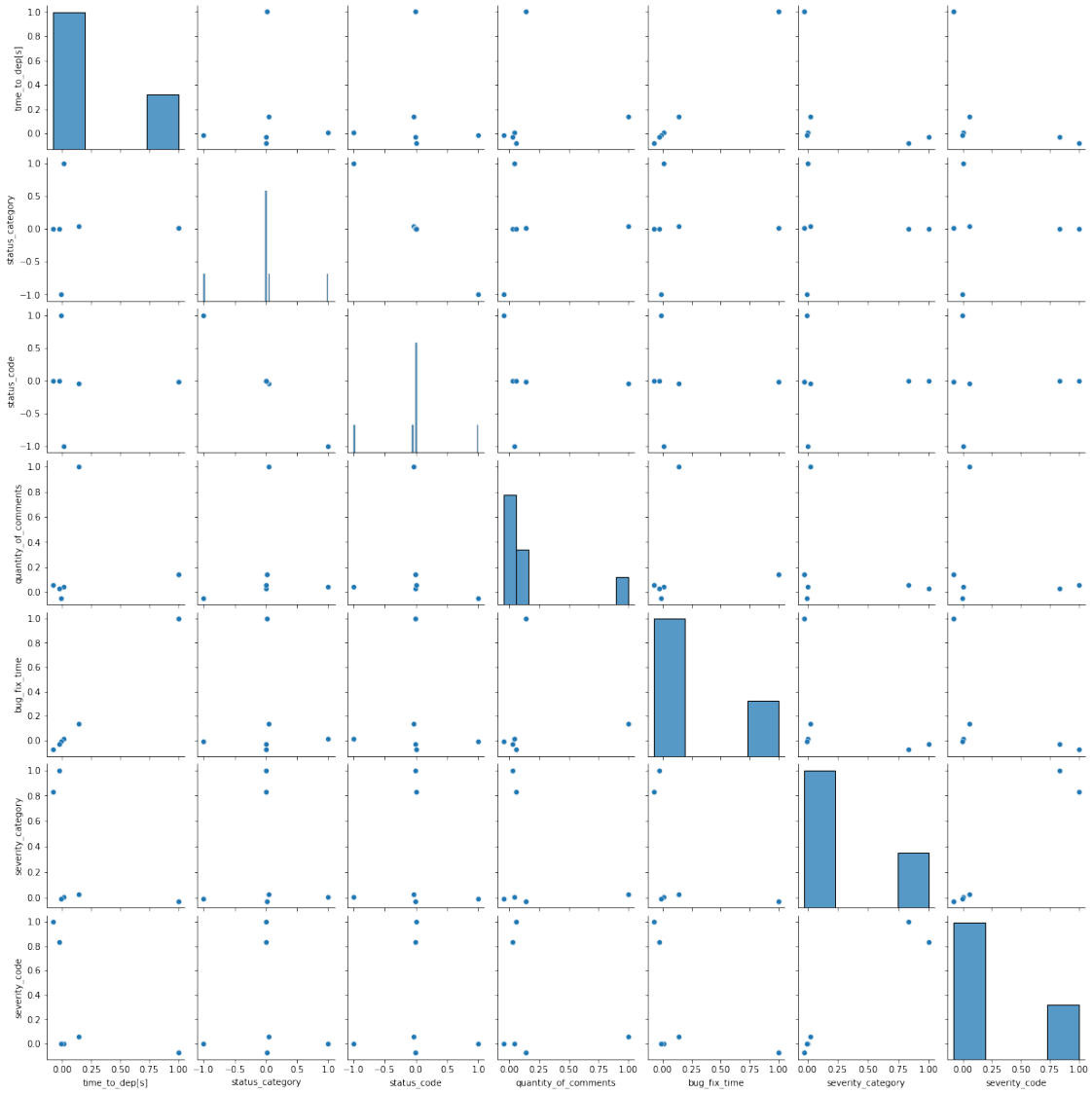
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]: <AxesSubplot:>



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]:
```
<seaborn.axisgrid.PairGrid at 0x266af73b9a0>
```



[ ]:
```
gcbrd.head()
```

[ ]:
```
   time_to_dep[s]  status_category  status_code  quantity_of_comments  \
0      10800000.0                1            4                     9
1       2505600.0                1            4                     5
```

```
2        518400.0                1            4                     5
3             0.0                1            4                     6
4       9936000.0                1            4                     5

   bug_fix_time  severity_category  severity_code
0           125                  0              2
1            29                  0              2
2             6                  0              2
3             0                  0              2
4           115                  0              2
```

[ ]:
```
X = gcbrd.iloc[:,:-1].values #rows and then columns in brackets
Y = gcbrd.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[1.0800e+07, 1.0000e+00, 4.0000e+00, 9.0000e+00, 1.2500e+02,
        0.0000e+00],
       [2.5056e+06, 1.0000e+00, 4.0000e+00, 5.0000e+00, 2.9000e+01,
        0.0000e+00],
       [5.1840e+05, 1.0000e+00, 4.0000e+00, 5.0000e+00, 6.0000e+00,
        0.0000e+00],
       …,
       [3.1968e+06, 1.0000e+00, 4.0000e+00, 1.7000e+01, 3.7000e+01,
        0.0000e+00],
       [1.7280e+05, 1.0000e+00, 4.0000e+00, 4.0000e+00, 2.0000e+00,
        0.0000e+00],
       [1.7280e+06, 1.0000e+00, 4.0000e+00, 9.0000e+00, 2.0000e+01,
        0.0000e+00]])
```

[ ]:
```
Y
```

[ ]:
```
array([2, 2, 2, …, 2, 2, 2], dtype=int64)
```

1.4.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
#
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=42)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
        # assignment which one we should accept from these
```

```
LinearRegression()
R2 score is:  0.6843366062851794
Mean Absolute error is:  0.14372935400170167
Mean Squared Error:  0.2115604173818574
---------------------------------------------------
LogisticRegression()
R2 score is:  -0.14746024808594949
Mean Absolute error is:  0.25400801603206413
Mean Squared Error:  0.7690380761523046
---------------------------------------------------
GaussianNB()
R2 score is:  -0.10260838171125441
Mean Absolute error is:  0.25701402805611223
Mean Squared Error:  0.7389779559118237
---------------------------------------------------
RandomForestClassifier()
R2 score is:  0.945430229244121
Mean Absolute error is:  0.008517034068136272
Mean Squared Error:  0.03657314629258517
---------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
```

```
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
--------------------------------------------------
SVR()
R2 score is:  -0.0267209009291951
Mean Absolute error is:  0.32418689291925057
Mean Squared Error:  0.6881174905300995
--------------------------------------------------
```

1.4.3 EDA On Gnome_bug_report_data

[ ]:
```
gnbrd.head()
```

[ ]:
```
             bug_id creation_date component_name product_name  \
0   NAUTILUS-438485    2007-05-15        general     NAUTILUS
1  EVOLUTION-231772    2002-10-05         Mailer    EVOLUTION
2        GLIB-59544    2001-08-25        general         GLIB
3         GGV-90943    2002-08-16        general          GGV
4        GDM-457958    2007-07-18        general          GDM

                                  short_description  \
0  crash in File Browser: -Navegar un directorio …
1                           Crash: closing evolution
2                 Provide closure support for GMain
3                                  Ui Isuue with ggv
4                           Hardcodes /sbin/nologin

                                   long_description  \
0  Version: 2.18.1\n\nWhat were you doing when th…
1  Package: Evolution\nPriority: Normal\nVersion:…
2  The GMain sources should be able to be connect…
3  The spin button on left top of the main window…
4  Hi\n\ngdm 2.19 series hardcode the nologin loc…

               assignee_name     reporter_name resolution_category  \
0             nautilus-maint   ricardo.ribalda               fixed
1  evolution-mail-maintainers           cgoheen               fixed
2                      gtkdev           otaylor               fixed
3                        jaka  satyajit.kanungo               fixed
4                   gdm-maint              lool               fixed

   resolution_code status_category  status_code update_date  \
0                1        resolved            4  2007-10-21
1                1        resolved            4  2013-09-10
2                1        resolved            4  2011-02-18
3                1        resolved            4  2004-12-22
4                1        resolved            4  2007-07-30
```

```
   quantity_of_votes  quantity_of_comments resolution_date  bug_fix_time  \
0                  0                     8      2007-10-21           159
1                  0                     6      2002-10-11             6
2                  0                     2      2001-09-07            13
3                  0                     4      2002-08-20             4
4                  0                     3      2007-07-30            12

  severity_category  severity_code
0          critical              5
1             major              4
2            normal              2
3            normal              2
4             major              4
```

[ ]:
```
gnbrd.tail()
```

[ ]:
```
                          bug_id creation_date              component_name  \
7808  EVOLUTION-DATA-SERVER-274212    2005-03-30                    Calendar
7809              GSTREAMER-515697    2008-02-11             gst-plugins-good
7810                    GTK+-84955    2002-06-11                Widget: Other
7811                    DIA-317980    2005-10-05                       shapes
7812               EPIPHANY-126397    2003-11-06  [obsolete] Backend:Mozilla

               product_name  \
7808  EVOLUTION-DATA-SERVER
7809              GSTREAMER
7810                   GTK+
7811                    DIA
7812               EPIPHANY

                                    short_description  \
7808         Crash while adding Fedora Release calendar
7809  [multifile] Several memory leaks exposed by un…
7810                    Gtkcombo popup window wrong size
7811         Breaking shape compatibility may break Dia
7812                          Crash while in background

                                     long_description     assignee_name  \
7808  Distribution: Fedora Core release 3 (Heidelber…  evolution-triage
7809  Hi\nthe attached patch fixes several memory le…    gstreamer-bugs
7810  When the horizontal scrollbar is showed in a c…          gtk-bugs
7811  Steps to reproduce:\n1.  use (possibly messed …              hans
7812  Distribution: Red Hat Linux release 9 (Shrike)…          mpgritti

     reporter_name resolution_category  resolution_code status_category  \
7808   e.rehrmann               fixed                1        resolved
```

```
7809         slomo               fixed                 1         resolved
7810       mpgritti              fixed                 1         resolved
7811            gab              fixed                 1         resolved
7812          aaron              fixed                 1         resolved

      status_code update_date  quantity_of_votes  quantity_of_comments  \
7808            4  2005-09-28                  0                     5
7809            4  2008-02-11                  0                     5
7810            4  2011-02-04                  0                     6
7811            4  2008-05-18                  0                     7
7812            4  2004-12-22                  0                     2

     resolution_date  bug_fix_time severity_category  severity_code
7808      2005-08-23           146            normal              2
7809      2008-02-11             0           blocker              6
7810      2002-10-10           121            normal              2
7811      2008-05-18           956          critical              5
7812      2003-11-06             0            normal              2
```

[ ]:
```
gnbrd.describe()
```

[ ]:
```
       resolution_code  status_code  quantity_of_votes  quantity_of_comments  \
count           7813.0       7813.0             7813.0           7813.000000
mean               1.0          4.0                0.0              6.097530
std                0.0          0.0                0.0             14.669013
min                1.0          4.0                0.0              0.000000
25%                1.0          4.0                0.0              2.000000
50%                1.0          4.0                0.0              4.000000
75%                1.0          4.0                0.0              7.000000
max                1.0          4.0                0.0            810.000000

       bug_fix_time  severity_code
count   7813.000000     7813.00000
mean     203.921157        2.56534
std      475.675999        1.22160
min       -1.000000        1.00000
25%        2.000000        2.00000
50%       24.000000        2.00000
75%      164.000000        2.00000
max     5577.000000        6.00000
```

[ ]:
```
gnbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7813 entries, 0 to 7812
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
```

```
---  ------                 --------------  -----
 0   bug_id                 7813 non-null   object
 1   creation_date          7813 non-null   object
 2   component_name         7813 non-null   object
 3   product_name           7813 non-null   object
 4   short_description      7792 non-null   object
 5   long_description       7784 non-null   object
 6   assignee_name          7813 non-null   object
 7   reporter_name          7813 non-null   object
 8   resolution_category    7813 non-null   object
 9   resolution_code        7813 non-null   int64
 10  status_category        7813 non-null   object
 11  status_code            7813 non-null   int64
 12  update_date            7813 non-null   object
 13  quantity_of_votes      7813 non-null   int64
 14  quantity_of_comments   7813 non-null   int64
 15  resolution_date        7813 non-null   object
 16  bug_fix_time           7813 non-null   int64
 17  severity_category      7813 non-null   object
 18  severity_code          7813 non-null   int64
dtypes: int64(6), object(13)
memory usage: 1.1+ MB
```

[ ]:
```
gnbrd['status_category'].unique()
```

[ ]:
```
array(['resolved'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status_category we can encode it as it is two categories only
# We are encoding it
gnbrd['status_category'] = gnbrd['status_category'].replace("resolved", 1)
```

[ ]:
```
gnbrd['severity_category'].unique()
```

[ ]:
```
array(['critical', 'major', 'normal', 'trivial', 'blocker', 'minor'],
      dtype=object)
```

[ ]:
```
# As we have another column severity_category and we can perform encoding on it␣
 ↪as well
gnbrd['severity_category'] = gnbrd['severity_category'].replace("normal", 0)
gnbrd['severity_category'] = gnbrd['severity_category'].replace("blocker", 1)
gnbrd['severity_category'] = gnbrd['severity_category'].replace("trivial", 2)
gnbrd['severity_category'] = gnbrd['severity_category'].replace("minor", 3)
gnbrd['severity_category'] = gnbrd['severity_category'].replace("major", 4)
gnbrd['severity_category'] = gnbrd['severity_category'].replace("critical", 5)
```

[ ]:
```
gnbrd.head(15)
```

[ ]:

	bug_id	creation_date	component_name	\
0	NAUTILUS-438485	2007-05-15	general	
1	EVOLUTION-231772	2002-10-05	Mailer	
2	GLIB-59544	2001-08-25	general	
3	GGV-90943	2002-08-16	general	
4	GDM-457958	2007-07-18	general	
5	EVOLUTION-218273	2002-01-06	Contacts	
6	PKG-CONFIG-142952	2004-05-22	general	
7	GUPPI-67853	2002-01-02	General	
8	GSTREAMER-637300	2010-12-15	gstreamer (core)	
9	GNOME-KEYRING-460902	2007-07-27	prompting	
10	SOUND-JUICER-517996	2008-02-22	docs	
11	EVOLUTION-207612	2001-08-19	Mailer	
12	GTK+-476823	2007-09-14	[obsolete] Backend: directfb	
13	WEBSITE-356286	2006-09-16	blogs.gnome.org	
14	SAWFISH-104778	2003-01-30	General	

	product_name	short_description	\
0	NAUTILUS	crash in File Browser: -Navegar un directorio …	
1	EVOLUTION	Crash: closing evolution	
2	GLIB	Provide closure support for GMain	
3	GGV	Ui Isuue with ggv	
4	GDM	Hardcodes /sbin/nologin	
5	EVOLUTION	scrollbar arrow widget should move the contact…	
6	PKG-CONFIG	Reorders the Libs: content	
7	GUPPI	Seg Fault when creating graph	
8	GSTREAMER	[API] request pad based on caps	
9	GNOME-KEYRING	keyring dialog is not modal	
10	SOUND-JUICER	Document keyboard shortcuts for editing	
11	EVOLUTION	Mailer crashes on check	
12	GTK+	Text coming as black rectangles	
13	WEBSITE	Urls with '+' sign like launchpad ones are wro…	
14	SAWFISH	sawfish won't register with gnome2	

	long_description	\
0	Version: 2.18.1\n\nWhat were you doing when th…	
1	Package: Evolution\nPriority: Normal\nVersion:…	
2	The GMain sources should be able to be connect…	
3	The spin button on left top of the main window…	
4	Hi\n\ngdm 2.19 series hardcode the nologin loc…	
5	Description of Problem:\nIf you have many cont…	
6	% cat /opt/gnome/lib/pkgconfig/evolution-sharp…	
7	Package: Guppi\nSeverity: normal\nVersion: 0.4…	
8	We should add new API to core to make requesti…	
9	The bug has been opened on https://bugs.launch…	
10	Sound-juicer should support keyboard shortcuts…	
11	Package: Evolution\nPriority: Blocker\nVersion…	

```
12  Please describe the problem:\nThe places where…
13  Please describe the problem:\nWhen you enter s…
14  Package: sawfish\nSeverity: major\nVersion: 1…

                assignee_name     reporter_name resolution_category  \
0               nautilus-maint   ricardo.ribalda               fixed
1   evolution-mail-maintainers           cgoheen               fixed
2                       gtkdev           otaylor               fixed
3                         jaka  satyajit.kanungo               fixed
4                   gdm-maint              lool               fixed
5                  sdevashish          jstrand1               fixed
6                      tfheen  tommi.komulainen               fixed
7                         jon       james.ogley               fixed
8               gstreamer-bugs             t.i.m               fixed
9          gnome-keyring-maint            seb128               fixed
10       gnome-user-docs-maint          jswitzer               fixed
11            evolution-triage              vwb2               fixed
12                   gtk-bugs       renymathara               fixed
13                      jdub         nbenitezl               fixed
14                       jsh           ctwardy               fixed

    resolution_code  status_category  status_code update_date  \
0                 1                 1            4  2007-10-21
1                 1                 1            4  2013-09-10
2                 1                 1            4  2011-02-18
3                 1                 1            4  2004-12-22
4                 1                 1            4  2007-07-30
5                 1                 1            4  2013-09-13
6                 1                 1            4  2008-04-28
7                 1                 1            4  2004-12-22
8                 1                 1            4  2011-01-05
9                 1                 1            4  2007-07-27
10                1                 1            4  2010-01-27
11                1                 1            4  2001-09-21
12                1                 1            4  2007-12-13
13                1                 1            4  2007-06-09
14                1                 1            4  2009-08-16

    quantity_of_votes  quantity_of_comments resolution_date  bug_fix_time  \
0                   0                     8      2007-10-21           159
1                   0                     6      2002-10-11             6
2                   0                     2      2001-09-07            13
3                   0                     4      2002-08-20             4
4                   0                     3      2007-07-30            12
5                   0                    14      2006-10-16          1744
6                   0                     3      2008-04-28          1437
7                   0                     2      2002-02-08            37
```

8	0	14	2011-01-05	21
9	0	2	2007-07-27	0
10	0	10	2010-01-27	705
11	0	3	2001-09-21	33
12	0	9	2007-12-13	90
13	0	4	2007-06-09	266
14	0	5	2003-08-15	197

	severity_category	severity_code
0	5	5
1	4	4
2	0	2
3	0	2
4	4	4
5	2	1
6	0	2
7	0	2
8	1	6
9	0	2
10	3	2
11	1	6
12	0	2
13	0	2
14	0	2

[ ]:
```
gnbrd['component_name'].unique()
```

[ ]:
```
array(['general', 'Mailer', 'Contacts', 'General', 'gstreamer (core)',
       'prompting', 'docs', '[obsolete] Backend: directfb',
       'blogs.gnome.org', 'gphoto backend', 'panel', 'gst-plugins-good',
       'Meta Contacts', 'Extension Library', 'Calendar',
       '[obsolete] bug-buddy parsing', 'Rendering', 'libgnomeui',
       'File and Folder Operations', 'win32', 'Plugins', 'interface',
       'clock', 'gst-libav', 'gst-plugins', 'Shell',
       'Widget: GtkComboBox', 'drivers', 'Misc', 'email', 'Trash',
       'User interface', '[obsolete] Backend:Mozilla',
       '[obsolete] Builds', 'nmcli', 'Do Not Use', 'Monitoring (inotify)',
       'Widget: Other', '[obsolete] Backgrounds Emblems and Themes',
       'baobab', 'Preferences', 'Printing', 'gst-plugins-bad', 'Tasks',
       'gtkvts', 'gnome-session', 'Interface', 'DAAP', 'Bookmarks',
       'Keyboard', 'gio', 'Mozilla interaction', 'Removable Media',
       'gtcd', 'mouse', 'Backend: Win32',
       '[obsolete] Preferred applications', 'menu applet',
       'import/export Text', 'gregex', 'demos', 'module sets',
       'Main System', 'GStreamer backend', 'application', 'filediff',
       'roboradio', 'Documentation', '[obsolete] hal volume monitor',
       'File Search Interface', 'HTTP Transport', 'User Documentation',
```

'magnifier', 'Code', 'network-admin',
'Passwords, Cookies, & Certificates', 'gail',
'Grapevine Client Library', 'gnibbles', 'users-admin', 'build',
'Movie player', 'modemlights', 'Metadata', 'GUI', 'Bindings',
'Module: (other)', 'doxywizard', 'ETable', 'widget support',
'livecd', 'gconf', 'ETree', 'libbonoboui', 'user interface',
'Man Pages', 'Views: Icon View', 'gtk', 'widgets', '.General',
'Build', 'Thumbnails', 'Framework', 'collection', 'Charting',
'[obsolete] Sidebar Panel: Tree', 'documentation', 'boot-admin',
'Views: All', 'gst-python', 'User Interface', 'Sheet Objects',
'reference documentation', 'YouTube plugin', 'gst-plugins-base',
'html-editor-control', 'Gimp-Python', 'Mathml', 'javabridge',
'Connector', 'gweather', '[obsolete] settings-daemon',
'notification area', '[obsolete] Views: Web Page', 'core',
'[obsolete] Assistive Technology Preferences', 'charpick',
'Device - USB Mass Storage', 'Widget: GtkFileChooser',
'File operations', 'Other Extensions', 'Sound',
'plugins: editor: scintilla', 'atk',
'keyboard indicator (gswitchit)', 'Background', 'Other', 'API',
'design', 'bug data', 'Preferences Dialog', 'Editing',
'Device - iPod', 'clearlooks', 'build_system', 'Views: List View',
'gnome', 'www.gimp.org', 'Importing', 'VPN: openvpn',
'Widget: GtkTreeView', 'gnect', 'Importers', 'VPN: pptp',
'libgnome-desktop', 'EText', 'cdplayer', 'performance',
'evolution', 'logview', 'gnome-terminal', 'Backend: X11',
'gst-plugins-ugly', 'dont know', 'libbonobomm', 'gsm',
'Mango (obsolete)', 'print preview', 'Tags', 'wizard',
'plugins: debug-manager', 'client', 'Playback', 'daemon',
'Windows Installer', 'user-guide', 'libbrasero-media',
'trash applet', 'gnome-power-manager', 'Reports', 'gobject',
'GalView', 'exports', 'window list', 'shell', 'Accessibility',
'Internationalisation', 'guadec.org', 'layout',
'import/export MS Excel (tm)', 'miscellaneous',
'[obsolete] Window preferences', 'LibGlade', 'Search Daemon',
'gsearchtool', 'window selector', 'Widget: GtkTextView',
'theme-highcontrastinverse', 'CDDBSlave2', 'gcalc',
'[obsolete] fonts:///', 'Tools', 'PDF backend', 'gmenu',
'tasklist', 'burn vfs-method', 'gst-rtsp-server',
'Widget: GtkIconView', 'magnifier-utility', 'gnotravex', 'libslab',
'trash backend', 'Blog Applet', 'chronojump',
'[obsolete] Keybinding', 'Other Preferences',
'plugins: language-support-c-cpp-java', 'speech', 'PTLIB',
'Desktop', 'codegen', 'UI', '[obsolete] Sidebar Panel: (Other)',
'XSLT', 'gdict', 'Browser plugin (obsolete)', 'Register',
'metatheme-edit', 'Podcast', 'Downloads', 'mahjongg', 'Sidebar',
'babl', 'mixer', 'gdict-applet', 'gcharmap', 'Gnome-CD',
'www.gnome.org', '[obsolete] GIO', 'gnome-about', 'file-chooser',

```
'Main Window', 'multiload', 'Canvas (libccc)',
'Programmatic interfaces', 'programs', 'filter all layers',
'libgnomecanvasmm', '[obsolete] nautilus-media', 'Browsing',
'mini-commander', '[obsolete] File types and programs', 'Engine',
'authorizations tool', 'Disk Mounter (drivemount)', 'toolbar',
'[obsolete] Sidebar', 'Website', 'glchess', 'screenshot', 'OPAL',
'video menu', 'items', 'Navigation', 'I18N', 'deskguide',
'gnome-cups-manager', 'Analytics', 'Backend: Quartz',
'Online Accounts', 'main', 'braille', 'Chat', 'applets', 'objects',
'Internationalization (i18n)', 'Script-Fu', 'power', 'updates',
'Search Tool', 'Network', 'keyring files', 'PostgreSQL provider',
'strings', '[obsolete] Views: Other', 'magnification', 'bonobo',
'Compilation', 'Message dialogs', 'core application',
'Class: UIManager / Actions', 'idetool', 'gnomoradio', 'hangul',
'Mouse', 'theme-highcontrast', 'Feeds', 'invest-applet',
'art.gnome.org (obsolete)', 'Disks UI', 'show-desktop-button',
'help.gnome.org', 'Import', 'Module: ssh/sftp', 'iagno',
'[obsolete] stock-icons', 'geyes', "Søren's compositor", 'capplet',
'MIME data', 'MIME and file/program mapping', 'Module: vfolder',
'Windows', 'Thumbnailer', 'battery', 'Plugins (other)', 'Default',
'idl-compiler', 'cpufreq', 'other', 'Simple bug guide', 'srcore',
'theme-mist', 'gnome-volume-control', 'Input Methods',
'Gnome-Sound-Recorder', 'Bindings Core', 'fonts', 'Markup backend',
'image viewer', 'Controls', 'atkbridge', 'boogle', 'tray',
'ftp backend', 'libbonobo', "Generally bug'd", 'cairo',
'cd-burner', 'Code Generator', 'shapes', 'packages',
'Device - MTP', 'playback', 'libseed', 'Index Daemon',
'Async operations', 'PS', 'gen_util', '[obsolete] Appearance',
'workspace switcher', 'Window Manager', 'BugBuddyBugs',
'[obsolete] Typing break', 'gnome-scan', 'composer', 'Themes',
'Group Chat', 'atsui', 'Parsing', 'gdk', 'i18n', 'shares-admin',
'[obsolete] theme-manager', 'plugins', 'Import - QIF', 'gnome-vfs',
'IP and DNS config', 'Semantic Analyzer', 'api', 'applet',
'Config Tool', 'Widget: GtkNotebook', 'Module: smb', 'Profile',
'conduit: memo_file', 'Loader', '[obsolete] about-me', 'freecell',
'gnomeapplet', 'Syntax files', 'Widget: GtkMenu',
'import/export HTML', 'gyrus-admin', 'xine-lib backend',
'Graphing / Charting', 'unknown', 'extensions', 'dialog',
'libgnome-menu', 'operations', 'editor', 'Debugging', 'Server',
'solaris', 'User Interface General', 'build utils', 'gst-omx',
'Widget: GtkPopover', 'GLX', 'gnome-sudoku',
'[obsolete] multihead', 'ripping', 'Power', 'glines', 'Tabs',
'[obsolete] Help System', 'libical', 'Display', 'Autotranslation',
'gnobots2', '[obsolete] Visual Design', "Iain's compositor",
'introspection', 'VU-meter', 'Smart Playlists', 'menu', 'bindings',
'gnome-power-statistics', 'registry', 'Grapevine Daemon',
'Enabling and Disabling', 'Keyboardability', 'themes',
```

'quadrapassel', 'locations', 'gsettings', 'Glib', 'Module: http',
'gpilotd', 'stickynotes', 'select-stylesheet', 'highcontrast',
'Backend', 'fish', 'products and taxonomy', 'import/export other',
'dataproviders', 'indic', 'EReflow', 'GAL Miscellaneous',
'documentation and translation', 'auto-scroller', 'orb-cpp',
'xalf', 'frontend', 'gtali', 'http backend', 'gstreamer-vaapi',
'ECategories', 'telepathy', 'settings', 'Miscellaneous / EWS Core',
'plugins: document-manager', 'Import - OFX', 'services-admin',
'Internet Radio', 'EWMH specification', 'libgnome', 'doc', 'www',
'mailcheck', 'player', 'xml2po', 'libanjuta', 'archive-generator',
'Composer', 'swell-foop', 'libgnome-scan', 'EFont', 'TreeView',
'process list', 'font-installer', 'l10n.gnome.org', 'adblock',
'install', 'YouTube service', '[obsolete] Keyboard Accessibility',
'libgbf', 'plugins: search', 'Printers', 'Postscript backend',
'Contacts (Global Address List)', 'error-viewer',
'Class: GtkRecent', 'bsd', 'Contact List', 'Cut Copy Paste Undo',
'Applet / Search Bar', 'playing', 'mail.gnome.org', 'zvt',
'Bonobo component', 'keyboard-accessibility (accessx-status)',
'gitg', 'ModemManager', 'cdda backend', 'Web Pages', 'Delegates',
'Section: Desktop Integration', 'Class: GtkBuilder', 'gui',
'plugins: project-manager', '[obsolete] screensaver', 'profiler',
'Build and Packaging', 'libgpa', 'Module: (compression/archiving)',
'Info Pages', 'Module: ftp', 'thai', 'gnome-power-preferences',
'Client library', 'libpanel-applet', 'gdialog', 'nm-applet',
'Widget: GtkToolbar', 'Help', 'time-admin', 'libgnomeprintuimm',
'Red Hat packages', 'Account', 'Sybase provider', 'gnome-help',
'Section: Windows', 'MySQL provider', 'totem', 'server side',
'English', 'media profiles', 'Gtk2', 'help-browser',
'gst-editing-services', 'Adwaita GTK3 theme', 'publisher', 'io',
'UOA', 'Scripts', 'Volume and drive handling', 'pager',
'URI handling', 'tp-aw', 'gthread', 'conversations', 'Jabber',
'Widget: GtkLabel', 'wireless-applet', 'DocBook', 'gtkhtml2',
'Test suite', 'media-keys', 'libvaladoc', 'database',
'svn.gnome.org (obsolete)', '[obsolete] BugBuddyBugs', 'events',
'smb', 'usability', 'bse', 'extensions.gnome.org', 'Installation',
's-t-b', 'object', '[obsolete] obexftp backend', 'sftp backend',
'editor widget', 'installer', 'libfolks', 'Memos', 'recent-files',
'Parse engine', 'Russian', 'theme-smokeyblue', 'Addressbook',
'Parser', 'Backend: Broadway', 'message-tray', 'page-info',
'Applet', 'GUI Expression Entry Widget', 'Backend - SQL',
'libgimp', 'developer.gnome.org', 'macos', 'Chat stack',
'gst-plugins-gl', 'gst-universe', 'plugins: glade', 'Docs',
'build infrastructure', 'Module: file', 'metadata',
'Developer Documentation', 'PDF', 'BDB backend', 'smart-bookmarks',
'Section: Visual Design', 'test', 'Wi-Fi', 'everything',
'MS OLE2 & Properties', 'plugins: symbol-db', 'developer-kit',
'sysadmin-guide', 'ldtprecord', 'libcryptui', 'gnomine',

```
       'obsolete', 'gfloppy', 'misc', 'wiki.gnome.org',
       'nm-connection-editor', 'Music and Sound Effects', 'LDAP provider',
       'signals', 'webdav backend', 'ldtp', 'Telepathy', 'Build system',
       'network-indicator', 'buildsystem', 'linux', 'theme-glider',
       'translation teams', 'MacOS', 'beast-gtk', 'Tokenizer',
       '[obsolete] Sound', 'gw', 'conduit system', 'dns-sd backend',
       'Timeline', 'download mirrors', 'overview', 'Game Engine',
       'Drawing', 'XML provider', 'VPN: vpnc', 'flegita'], dtype=object)
```

[ ]:
```
gnbrd['product_name'].unique()
```

[ ]:
```
array(['NAUTILUS', 'EVOLUTION', 'GLIB', 'GGV', 'GDM', 'PKG-CONFIG',
       'GUPPI', 'GSTREAMER', 'GNOME-KEYRING', 'SOUND-JUICER', 'GTK+',
       'WEBSITE', 'SAWFISH', 'GVFS', 'GNOME-CORE', 'PANGO', 'EMPATHY',
       'TOTEM', 'GLIBMM', 'GALEON', 'BEAGLE', 'BUGZILLA.GNOME.ORG',
       'GTKHTML', 'GNOME-LIBS', 'DIA', 'GIMP', 'GNOME-PANEL',
       'GNOME-TERMINAL', 'GAZPACHO', 'GFTP', 'CHEESE',
       'EVOLUTION-DATA-SERVER', 'LIBGNOME', 'GNOME-SPEECH', 'BOUNTIES',
       'LIBGSF', 'GTHUMB', 'GEDIT', 'BRASERO', 'GOSSIP', 'GNOME-TODO',
       'GNOME-ROBOTS', 'EPIPHANY', 'NETWORKMANAGER', 'GNOME-PRINT',
       'GNOME-VFS', 'GNOME-POWER-MANAGER', 'SYSTEM-MONITOR', 'GTKMM',
       'GNOME-UTILS', 'TRACKER', 'RHYTHMBOX', 'GNOME-CONTROL-CENTER',
       'GTKVTS', 'GNOME-SESSION', 'LIBZVT', 'GNOMEICU', 'GARNOME',
       'PYGTK', 'GNUMERIC', 'GNOME-MEDIA', 'GNOME-SETTINGS-DAEMON',
       'EVINCE', 'GCONF-EDITOR', 'GNOME-APPLETS', 'LIBGNOMEUI', 'STRAW',
       'PAN', 'JHBUILD', 'GTETRINET', 'GNOME-SHELL', 'BLUEFISH', 'MELD',
       'GLADE', 'GNOMORADIO', 'EKIGA', 'GNOME-NETTOOL', 'LIBSOUP',
       'METACITY', 'GHEX', 'GNOME-LOGS', 'BONOBO-ACTIVATION_[WAS:_OAF]',
       'GNOPERNICUS', 'LIBXML++', 'EEL', 'CONGLOMERATE',
       'GNOME-SYSTEM-TOOLS', 'GJS', 'BUG-BUDDY', 'LIBRSVG', 'ATK',
       'GRAPEVINE', 'GNOME-GAMES-SUPERSEDED', 'GNOME-CALCULATOR',
       'DESKBAR-APPLET', 'XCHAT-GNOME', 'TOMBOY', 'PESSULUS', 'F-SPOT',
       'VALA', 'LIBWNCK', 'DOXYGEN', 'GNOME-COMMON', 'GDK-PIXBUF', 'VTE',
       'GNOME-SUBTITLES', 'ACCERCISER', 'GAL', 'GPARTED', 'GCONF',
       'BONOBO', 'SEAHORSE', 'LSR', 'YELP', 'LIBGNOMEDB', 'DOGTAIL',
       'GCOMPRIS', 'QUICK-LOUNGE-APPLET', 'BANSHEE', 'EOG', 'LAST-EXIT',
       'PYGOBJECT', 'DAMNED-LIES', 'DEVHELP', 'PLANNER', 'GUCHARMAP',
       'LASEM', 'AT-SPI', 'MUINE', 'EVOLUTION_EXCHANGE', 'DASHER',
       'LIBXSLT', 'ANJUTA', 'GLOM', 'VINO', 'FILE-ROLLER', 'FOLKS', 'GOK',
       'SHOTWELL', 'ORCA', 'GTK-ENGINES', 'GDESKLETS', 'GEDIT-PLUGINS',
       'GNOME-MAIN-MENU', 'GNOME-PYTHON', 'GIMP-WEB', 'LIBGTOP',
       'GNOME-DESKTOP', 'GNOME-CHESS', 'XMLSEC', 'GNOME-PYTHON-DESKTOP',
       'DRIVEL', 'GNOMEMM', 'LIBGNOME-KEYRING', 'SYSADMIN',
       'GNOME-BLUETOOTH', 'GNOME-SCREENSAVER', 'GEARY', 'GNOME-USER-DOCS',
       'GNUCASH', 'DEVILSPIE', 'GNOME-MENUS', 'SCAFFOLD', 'LIBGOFFICE',
       'JAVA-GNOME', 'LDTP', 'GNET', 'PYORBIT', 'MEDUSA',
       'GOBJECT-INTROSPECTION', 'GTK-DOC', 'BALSA', 'GNOME-THEMES',
```

```
'GNOME-COMMANDER', 'MARLIN', 'NAUTILUS-CD-BURNER', 'GDL',
'GPERFMETER', 'GNOME-MAG', 'LIBGWEATHER', 'GNOME-BLOG',
'CHRONOJUMP', 'SERPENTINE', 'GTK-VNC', 'GNOME-DB', 'VALENCIA',
'GMIME', 'BOOKWORM', 'ADWAITA-ICON-THEME', 'METATHEME', 'GNORPM',
'GEGL', 'INTLTOOL', 'CONDUIT', 'GNOME-DEVEL-DOCS', 'CRIAWIPS',
'LIBGLADE', 'GNOME-MOUNT', 'GIMP-GAP', 'GNOME-VFSMM', 'LIBGLADEMM',
'AT-POKE', 'GAMIN', 'GIMP-MANUAL', 'POLICYKIT-GNOME',
'GLADE-LEGACY', 'ACME', 'GNOME-KEYRING-MANAGER', 'LIBEGG',
'GTRANSLATOR', 'DRWRIGHT', 'EPIPHANY-EXTENSIONS', 'GPDF',
'LIBGNOMECANVAS', 'GNOME-CUPS-MANAGER', 'GNOTE',
'FILEMANAGER-ACTIONS', 'MEMPROF', 'MERGEANT', 'GRILO', 'LIBGDA',
'GTKSOURCEVIEW', 'GTKGLEXT', 'LIBGEE', 'LIBGTCPSOCKET', 'MLVIEW',
'SABAYON', 'BLAM', 'GIMMIE', 'GNOME-DISK-UTILITY', 'GCM',
'EVOLUTION-SCALIX', 'GNOME-MIME-DATA', 'ORBIT-CPP', 'GLIB-OPENSSL',
'NAUTILUS-SENDTO', 'BAKERY', 'BATTFINK', 'GNOME-SCHEDULE', 'GWGET',
'MUTTER', 'SEED', 'GNOME-SCAN', 'LIBIDL', 'GCONFMM',
'GSTREAMER-VAAPI', 'GNOME-MAHJONGG', 'LIBXML', 'GNOMERADIO',
'GNOME-PERL', 'GNOME-PILOT', 'NEMIVER', 'MCATALOG', 'ISTANBUL',
'PYPHANY', 'GNOME-BOXES', 'PASSEPARTOUT', 'GYRUS', 'HIPO', 'LINC',
'HAMSTER-APPLET', 'GNOME-BUILDER', 'GNOME-PHONE-MANAGER',
'CONTACT-LOOKUP-APPLET', 'LIBNOTIFY', 'GNOME-DEBUG',
'GNOME-DOC-UTILS', 'ZENITY', 'COGL', 'ESOUND', 'LIBGNOMEUIMM',
'GNOME-THEMES-EXTRAS', 'GNOME-THEMES-STANDARD', 'ORBIT',
'GNOME-BACKGROUNDS', 'GNOME-MUD', 'GNOME-VOLUME-MANAGER',
'LIBSIGC++', 'TOTEM-PL-PARSER', 'EVOLUTION-SHARP',
'GNOME-FILE-SELECTOR', 'OSTREE', 'SNOWY', 'XALF',
'GSETTINGS-DESKTOP-SCHEMAS', 'PRINTMAN', 'TASQUE', 'EVOLUTION-EWS',
'DASHBOARD', 'GOOBOX', 'PITIVI', 'GNOME-DOCUMENTS',
'GNOME-USER-SHARE', 'BIJIBEN', 'LIBGDATA', 'GNOME-BUILD',
'XSCREENSAVER', 'GNOME-ONLINE-ACCOUNTS', 'HIG', 'ALACARTE',
'VINAGRE', 'POSTR', 'GITG', 'EAZEL-TOOLS', 'EAZEL-HACKING',
'MOUSETWEAKS', 'GNOME-AUTOAR', 'GALF', 'LIBGOVIRT', 'LIBEPC',
'GIMP-TINY-FU', 'RESAPPLET', 'GNOME-CLOCKS', 'CLUTTERMM',
'YELP-XSL', 'GNOME-PYTHON-EXTRAS', 'VALADOC', 'GNOME-SOFTWARE',
'CALIFORNIA', 'PAPERBOX', 'GNOME-VFS-EXTRAS', 'BEAST', 'JAMBOREE',
'AISLERIOT', 'MAGICDEV', 'GENERAL', 'GNOME-FORMAT', 'LIBGDAMM',
'YARRR', 'GUIKACHU', 'DESKTOP-FILE-UTILS', 'GNOME-NETWORK',
'GNOME-SUDOKU', 'SWFDEC-GNOME', 'GNOME-NETSTATUS', 'BIGBOARD',
'LIBBTCTL', 'FIVE-OR-MORE', 'GAUPOL', 'TEST', 'RYGEL', 'HOTSSH',
'ONTV', 'REMOVABLE_MEDIA_MANAGER', 'GNOME-LIVE', 'CUPID',
'MONKEY-BUBBLE', 'GNOME-CALENDAR', 'PYSPI', 'NEWCOMERS-TUTORIAL',
'GUPNP-AV', 'LIBGNOMEKBD', 'TOUTDOUX'], dtype=object)
```

1.5 As other elements are not having a categorical

1.5.1 Now we will be removing null values

[ ]:
```
gnbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description      21
long_description       29
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
# percentage of missing values
gnbrd.isnull().sum() / gnbrd.shape[0] * 100
```

[ ]:
```
bug_id                  0.000000
creation_date           0.000000
component_name          0.000000
product_name            0.000000
short_description       0.268783
long_description        0.371176
assignee_name           0.000000
reporter_name           0.000000
resolution_category     0.000000
resolution_code         0.000000
status_category         0.000000
status_code             0.000000
update_date             0.000000
quantity_of_votes       0.000000
quantity_of_comments    0.000000
resolution_date         0.000000
bug_fix_time            0.000000
severity_category       0.000000
```

```
severity_code           0.000000
dtype: float64
```

[ ]:
```
gnbrd.dropna(inplace=True)
```

[ ]:
```
gnbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description        0
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
gnbrd.head()
```

[ ]:
```
             bug_id creation_date component_name product_name  \
0   NAUTILUS-438485    2007-05-15        general     NAUTILUS
1  EVOLUTION-231772    2002-10-05         Mailer    EVOLUTION
2        GLIB-59544    2001-08-25        general         GLIB
3         GGV-90943    2002-08-16        general          GGV
4        GDM-457958    2007-07-18        general          GDM

                                 short_description  \
0  crash in File Browser: -Navegar un directorio …
1                          Crash: closing evolution
2                 Provide closure support for GMain
3                                 Ui Isuue with ggv
4                          Hardcodes /sbin/nologin

                                  long_description  \
0  Version: 2.18.1\n\nWhat were you doing when th…
1  Package: Evolution\nPriority: Normal\nVersion:…
```

```
2  The GMain sources should be able to be connect…
3  The spin button on left top of the main window…
4  Hi\n\ngdm 2.19 series hardcode the nologin loc…

                assignee_name     reporter_name resolution_category  \
0              nautilus-maint   ricardo.ribalda               fixed
1  evolution-mail-maintainers           cgoheen               fixed
2                      gtkdev           otaylor               fixed
3                        jaka  satyajit.kanungo               fixed
4                   gdm-maint              lool               fixed

   resolution_code  status_category  status_code update_date  \
0                1                1            4  2007-10-21
1                1                1            4  2013-09-10
2                1                1            4  2011-02-18
3                1                1            4  2004-12-22
4                1                1            4  2007-07-30

   quantity_of_votes  quantity_of_comments resolution_date  bug_fix_time  \
0                  0                     8      2007-10-21           159
1                  0                     6      2002-10-11             6
2                  0                     2      2001-09-07            13
3                  0                     4      2002-08-20             4
4                  0                     3      2007-07-30            12

   severity_category  severity_code
0                  5              5
1                  4              4
2                  0              2
3                  0              2
4                  4              4
```

[ ]:
```
from datetime import time,date,datetime

gnbrd['update_date'] = pd.to_datetime(gnbrd['update_date'])
gnbrd['creation_date'] = pd.to_datetime(gnbrd['creation_date'])
gnbrd['resolution_date'] = pd.to_datetime(gnbrd['resolution_date'])
```

[ ]:
```
gnbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7763 entries, 0 to 7812
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                7763 non-null   object
 1   creation_date         7763 non-null   datetime64[ns]
```

```
 2   component_name        7763 non-null   object
 3   product_name          7763 non-null   object
 4   short_description     7763 non-null   object
 5   long_description      7763 non-null   object
 6   assignee_name         7763 non-null   object
 7   reporter_name         7763 non-null   object
 8   resolution_category   7763 non-null   object
 9   resolution_code       7763 non-null   int64
 10  status_category       7763 non-null   int64
 11  status_code           7763 non-null   int64
 12  update_date           7763 non-null   datetime64[ns]
 13  quantity_of_votes     7763 non-null   int64
 14  quantity_of_comments  7763 non-null   int64
 15  resolution_date       7763 non-null   datetime64[ns]
 16  bug_fix_time          7763 non-null   int64
 17  severity_category     7763 non-null   int64
 18  severity_code         7763 non-null   int64
dtypes: datetime64[ns](3), int64(8), object(8)
memory usage: 1.2+ MB
```

[ ]:
```
gnbrd.insert(0, "time_to_dep[s]", ((
    gnbrd['resolution_date']-gnbrd['creation_date']).astype('timedelta64[s]')),␣
 ↪True)
```

[ ]:
```
gnbrd.head()
```

[ ]:
```
   time_to_dep[s]           bug_id creation_date component_name product_name  \
0      13737600.0   NAUTILUS-438485    2007-05-15        general     NAUTILUS
1        518400.0  EVOLUTION-231772    2002-10-05         Mailer    EVOLUTION
2       1123200.0        GLIB-59544    2001-08-25        general         GLIB
3        345600.0         GGV-90943    2002-08-16        general          GGV
4       1036800.0        GDM-457958    2007-07-18        general          GDM

                                short_description  \
0  crash in File Browser: -Navegar un directorio …
1                          Crash: closing evolution
2                Provide closure support for GMain
3                                Ui Isuue with ggv
4                          Hardcodes /sbin/nologin

                                 long_description  \
0  Version: 2.18.1\n\nWhat were you doing when th…
1  Package: Evolution\nPriority: Normal\nVersion:…
2  The GMain sources should be able to be connect…
3  The spin button on left top of the main window…
4  Hi\n\ngdm 2.19 series hardcode the nologin loc…
```

```
                assignee_name     reporter_name resolution_category  \
0              nautilus-maint   ricardo.ribalda               fixed
1  evolution-mail-maintainers           cgoheen               fixed
2                       gtkdev           otaylor               fixed
3                         jaka  satyajit.kanungo               fixed
4                    gdm-maint              lool               fixed

   resolution_code  status_category  status_code update_date  \
0                1                1            4  2007-10-21
1                1                1            4  2013-09-10
2                1                1            4  2011-02-18
3                1                1            4  2004-12-22
4                1                1            4  2007-07-30

   quantity_of_votes  quantity_of_comments resolution_date  bug_fix_time  \
0                  0                     8      2007-10-21           159
1                  0                     6      2002-10-11             6
2                  0                     2      2001-09-07            13
3                  0                     4      2002-08-20             4
4                  0                     3      2007-07-30            12

   severity_category  severity_code
0                  5              5
1                  4              4
2                  0              2
3                  0              2
4                  4              4
```

[ ]:
```
# dropping extra columns
gnbrd.
 ↪drop(['bug_id','resolution_code','status_category','status_code','quantity_of_votes','creat
```

[ ]:
```
# now we are done with dropping values as well
gnbrd['bug_fix_time'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew         4.262647
kurtosis    23.079167
Name: bug_fix_time, dtype: float64
```

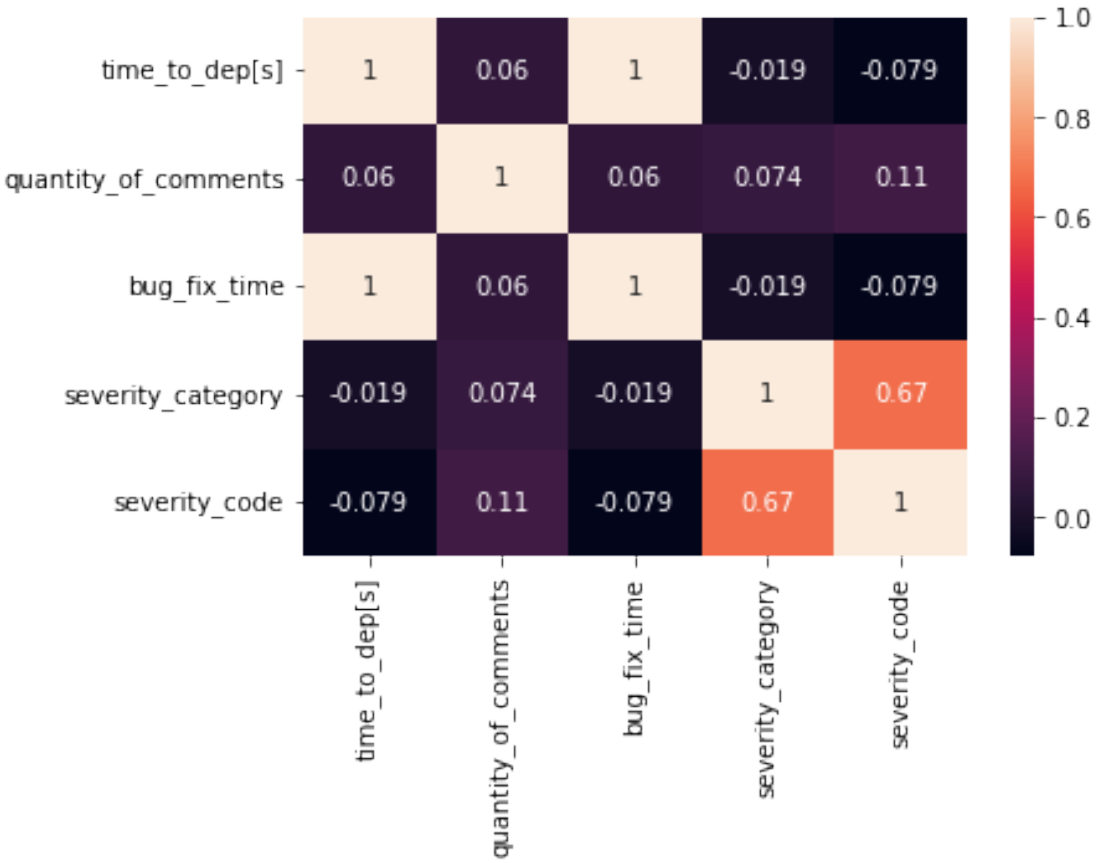
[ ]:
```
corr = gnbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  quantity_of_comments  bug_fix_time  \
time_to_dep[s]              1.000000              0.059621      1.000000
quantity_of_comments        0.059621              1.000000      0.059621
bug_fix_time                1.000000              0.059621      1.000000
severity_category          -0.019349              0.073711     -0.019349
```

```
severity_code              -0.079069              0.106376      -0.079069

                       severity_category  severity_code
time_to_dep[s]                -0.019349      -0.079069
quantity_of_comments           0.073711       0.106376
bug_fix_time                  -0.019349      -0.079069
severity_category              1.000000       0.670572
severity_code                  0.670572       1.000000
```

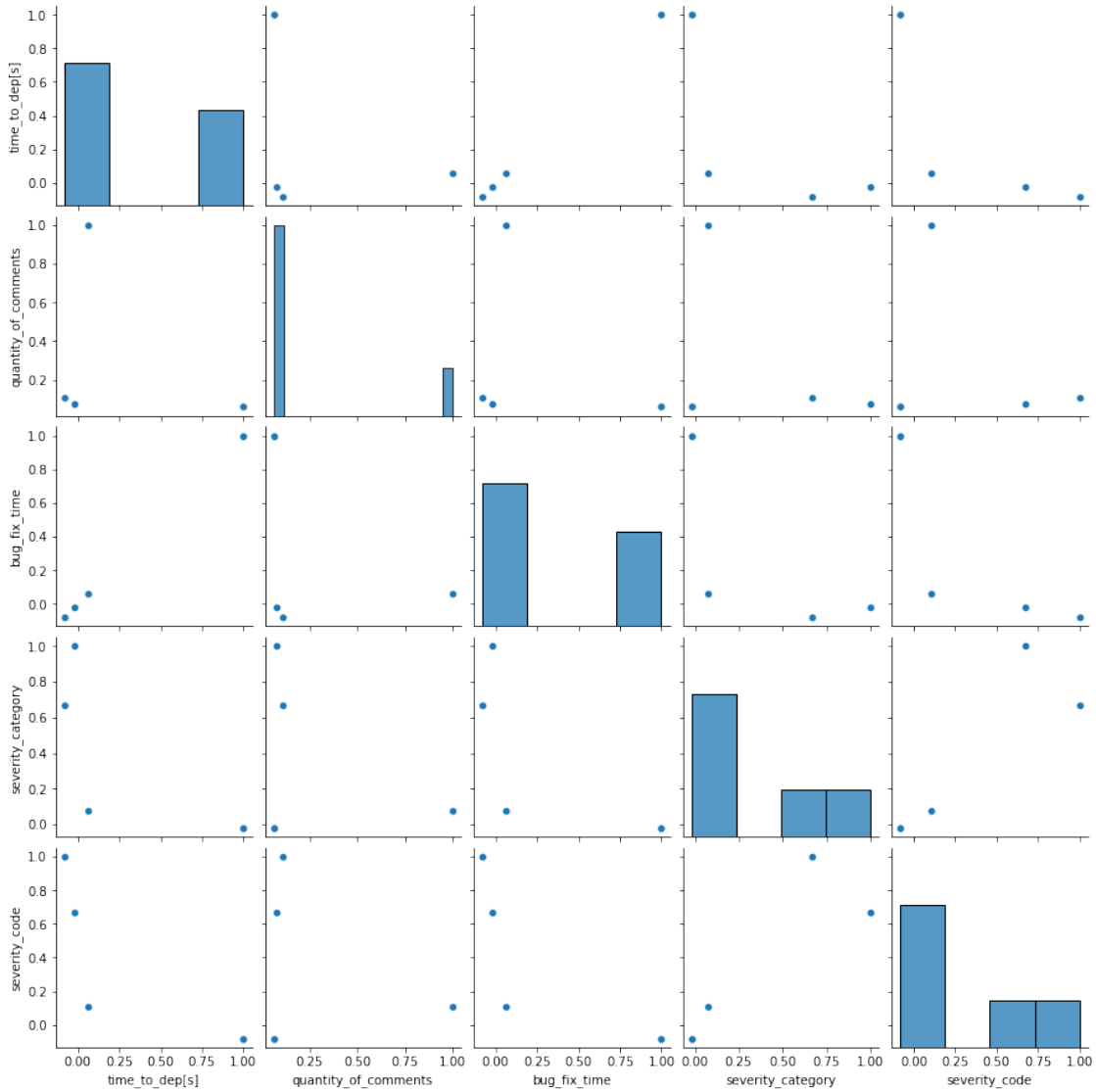
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]:
```
<AxesSubplot:>
```



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]:
```
<seaborn.axisgrid.PairGrid at 0x266b1907fd0>
```



[ ]:
```
gnbrd.head()
```

[ ]:
```
   time_to_dep[s]  quantity_of_comments  bug_fix_time  severity_category  \
0      13737600.0                     8           159                  5
1        518400.0                     6             6                  4
2       1123200.0                     2            13                  0
3        345600.0                     4             4                  0
4       1036800.0                     3            12                  4

   severity_code
0              5
1              4
2              2
```

```
3               2
4               4
```

[ ]:
```
X = gnbrd.iloc[:, :-1].values  # rows and then columns in brackets
Y = gnbrd.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[1.37376e+07, 8.00000e+00, 1.59000e+02, 5.00000e+00],
       [5.18400e+05, 6.00000e+00, 6.00000e+00, 4.00000e+00],
       [1.12320e+06, 2.00000e+00, 1.30000e+01, 0.00000e+00],
       …,
       [1.04544e+07, 6.00000e+00, 1.21000e+02, 0.00000e+00],
       [8.25984e+07, 7.00000e+00, 9.56000e+02, 5.00000e+00],
       [0.00000e+00, 2.00000e+00, 0.00000e+00, 0.00000e+00]])
```

[ ]:
```
Y
```

[ ]:
```
array([5, 4, 2, …, 2, 5, 2], dtype=int64)
```

1.5.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
#
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=42)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
```

```
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
        # assignment which one we should accept from these
```

```
LinearRegression()
R2 score is:  0.4046313377733384
Mean Absolute error is:  0.47469467693759915
Mean Squared Error:  0.8693569498555661
---------------------------------------------------
LogisticRegression()
R2 score is:  -0.18446544048621671
Mean Absolute error is:  0.609143593045718
Mean Squared Error:  1.7295556986477785
---------------------------------------------------
GaussianNB()
R2 score is:  -0.18446544048621671
Mean Absolute error is:  0.609143593045718
Mean Squared Error:  1.7295556986477785
---------------------------------------------------
RandomForestClassifier()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
---------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
---------------------------------------------------
SVR()
R2 score is:  -0.12004460608377299
Mean Absolute error is:  0.6686873428570365
Mean Squared Error:  1.63548843636729
---------------------------------------------------
```

1.5.3 EDA On Mozila_Bug_Report_Data

[ ]:
```
mbrd.head()
```

[ ]:
```
                    bug_id creation_date    component_name  \
0            BUGZILLA-294734    2005-05-18  Bugzilla-General
```

```
1   OTHER_APPLICATIONS-363323    2006-12-09     DOM Inspector
2  SUPPORT.MOZILLA.ORG-398246    2007-10-02          General
3  RELEASE_ENGINEERING-525991    2009-11-02          General
4   OTHER_APPLICATIONS-318859    2005-12-02        ChatZilla

          product_name                                  short_description  \
0              BUGZILLA                          Emergency 2.16.10 Release
1   OTHER_APPLICATIONS  DOM View is really inefficient with setting wh…
2  SUPPORT.MOZILLA.ORG   Add support for custom cookies and cache headers
3  RELEASE_ENGINEERING          Create Major Update from 3.0.15 to 3.5.5
4   OTHER_APPLICATIONS   DCC functionality in ChatZilla isn't functional.

                                  long_description        assignee_name  \
0  2.16.9 is broken -- many users can't enter bug…              mkanat
1  From comment in url:\n\nCurrent code:\nmenuite…             sdwilsh
2  Adding support for custom headers and cookie n…            morgamic
3                                              NaN               catlee
4  User-Agent:       Mozilla/5.0 (Macintosh U PPC…  gijskruitbosch+bugs

  reporter_name resolution_category  resolution_code status_category  \
0       mkanat               fixed               1        resolved
1      sdwilsh               fixed               1        resolved
2     morgamic               fixed               1        resolved
3       catlee               fixed               1        resolved
4       dafydd               fixed               1        resolved

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            4  2005-05-19                  0                    15
1            4  2011-06-01                  0                     8
2            4  2009-11-02                  0                    23
3            4  2013-08-12                  0                     7
4            4  2006-02-10                  0                    14

  resolution_date  bug_fix_time severity_category  severity_code
0      2005-05-19             1           blocker              6
1      2007-01-14            36            normal              2
2      2008-03-24           174           blocker              6
3      2009-11-10             8            normal              2
4      2005-12-07             5            normal              2
```

[ ]:
```
mbrd.tail()
```

[ ]:
```
                          bug_id creation_date    component_name  \
9994        WWW.MOZILLA.ORG-485595    2009-03-27           General
9995                    CORE-132278    2002-03-20             XPCOM
9996   FIREFOX_BUILD_SYSTEM-389793    2007-07-26           General
9997  MOZILLA_LOCALIZATIONS-402568    2007-11-05   fy-NL / Frisian
```

```
9998                  CORE-294989    2005-05-20  Spelling checker

             product_name  \
9994        WWW.MOZILLA.ORG
9995                  CORE
9996   FIREFOX_BUILD_SYSTEM
9997  MOZILLA_LOCALIZATIONS
9998                  CORE

                                    short_description  \
9994  Firefox 2.0. - 2.0.0.20 update/what's new page…
9995  |nsCOMPtr::operator&()| has outlived its usefu…
9996  Firefox build failed on OpenSolaris without --…
9997  language pack fails to install due to broken i…
9998  In  4 Warning: anonymous function does not alw…

                                     long_description      assignee_name  \
9994  User-Agent:       Mozilla/5.0 (Windows U Windo…              nobody
9995  We originally made |operator&| illegal to ease…                 scc
9996  gmake[6]: Entering directory `/export/home/mrb…  ginnchen+exoracle
9997  The files \n  l10n/fy-NL/browser/defines.inc\n…      fryskefirefox
9998  [Mozilla/5.0 (Windows U Win98 en-US rv:1.8b2) …            mnyromyr

                       reporter_name resolution_category  resolution_code  \
9994                       domthedude001               fixed                1
9995                                scc               fixed                1
9996                  ginnchen+exoracle               fixed                1
9997                            nthomas               fixed                1
9998  bugzillamozillaorg_serge_20140323               fixed                1

     status_category  status_code update_date  quantity_of_votes  \
9994        resolved            4  2012-08-23                  0
9995        resolved            4  2003-01-16                  0
9996        resolved            4  2018-03-02                  0
9997        resolved            4  2009-11-27                  0
9998        resolved            4  2005-06-24                  0

      quantity_of_comments resolution_date  bug_fix_time severity_category  \
9994                    15      2010-12-27           640           trivial
9995                    10      2002-03-28             8            normal
9996                    15      2007-08-05            10            normal
9997                     3      2009-11-27           753            normal
9998                     6      2005-06-24            35             minor

      severity_code
9994              1
9995              2
```

```
9996              2
9997              2
9998              2
```

[ ]:
```
mbrd.describe()
```

[ ]:
```
       resolution_code  status_code  quantity_of_votes  quantity_of_comments  \
count           9999.0  9999.000000        9999.000000           9999.000000
mean               1.0     4.004800           0.360536             12.769877
std                0.0     0.097872           2.862126             17.589481
min                1.0     4.000000           0.000000              1.000000
25%                1.0     4.000000           0.000000              4.000000
50%                1.0     4.000000           0.000000              8.000000
75%                1.0     4.000000           0.000000             15.000000
max                1.0     6.000000         101.000000            407.000000

       bug_fix_time  severity_code
count   9999.000000    9999.000000
mean     278.098510       2.380938
std      661.487424       1.026748
min        0.000000       1.000000
25%        3.000000       2.000000
50%       24.000000       2.000000
75%      188.000000       2.000000
max     7294.000000       6.000000
```

[ ]:
```
mbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9999 entries, 0 to 9998
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   bug_id               9999 non-null   object
 1   creation_date        9999 non-null   object
 2   component_name       9999 non-null   object
 3   product_name         9999 non-null   object
 4   short_description    9998 non-null   object
 5   long_description     9920 non-null   object
 6   assignee_name        9999 non-null   object
 7   reporter_name        9999 non-null   object
 8   resolution_category  9999 non-null   object
 9   resolution_code      9999 non-null   int64
 10  status_category      9999 non-null   object
 11  status_code          9999 non-null   int64
 12  update_date          9999 non-null   object
 13  quantity_of_votes    9999 non-null   int64
```

```
 14  quantity_of_comments  9999 non-null   int64
 15  resolution_date       9999 non-null   object
 16  bug_fix_time          9999 non-null   int64
 17  severity_category     9999 non-null   object
 18  severity_code         9999 non-null   int64
dtypes: int64(6), object(13)
memory usage: 1.4+ MB
```

[ ]:
```
mbrd['status_category'].unique()
```

[ ]:
```
array(['resolved', 'closed'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status_category we can encode it as it is two categories only
# We are encoding it
mbrd['status_category'] = mbrd['status_category'].replace("resolved", 1)
mbrd['status_category'] = mbrd['status_category'].replace("closed", 0)
```

[ ]:
```
mbrd['severity_category'].unique()
```

[ ]:
```
array(['blocker', 'normal', 'major', 'minor', 'critical', 'trivial'],
      dtype=object)
```

[ ]:
```
# As we have another column severity_category and we can perform encoding on it␣
 ↪as well
mbrd['severity_category'] = mbrd['severity_category'].replace("normal", 0)
mbrd['severity_category'] = mbrd['severity_category'].replace("blocker", 1)
mbrd['severity_category'] = mbrd['severity_category'].replace("trivial", 2)
mbrd['severity_category'] = mbrd['severity_category'].replace("minor", 3)
mbrd['severity_category'] = mbrd['severity_category'].replace("major", 4)
mbrd['severity_category'] = mbrd['severity_category'].replace("critical", 5)
```

[ ]:
```
mbrd.head(15)
```

[ ]:

	bug_id	creation_date	component_name	\
0	BUGZILLA-294734	2005-05-18	Bugzilla-General	
1	OTHER_APPLICATIONS-363323	2006-12-09	DOM Inspector	
2	SUPPORT.MOZILLA.ORG-398246	2007-10-02	General	
3	RELEASE_ENGINEERING-525991	2009-11-02	General	
4	OTHER_APPLICATIONS-318859	2005-12-02	ChatZilla	
5	DEVELOPER.MOZILLA.ORG-416840	2008-02-11	General	
6	MARKETPLACE_GRAVEYARD-977729	2014-02-27	Reviewer Tools	
7	THUNDERBIRD-339875	2006-05-31	Build Config	
8	FIREFOX-420556	2008-03-02	Theme	
9	MOZILLA_LOCALIZATIONS-415621	2008-02-04	mk / Macedonian	
10	CORE-254510	2004-08-05	Image Blocking	
11	CORE-204573	2003-05-05	HTML: Parser	

```
12                  THUNDERBIRD-412434    2008-01-15           General
13  OTHER_APPLICATIONS_GRAVEYARD-391445    2007-08-08      QA Companion
14          FIREFOX_OS_GRAVEYARD-942470    2013-11-23           General

                   product_name  \
0                      BUGZILLA
1             OTHER_APPLICATIONS
2            SUPPORT.MOZILLA.ORG
3            RELEASE_ENGINEERING
4             OTHER_APPLICATIONS
5          DEVELOPER.MOZILLA.ORG
6          MARKETPLACE_GRAVEYARD
7                    THUNDERBIRD
8                        FIREFOX
9          MOZILLA_LOCALIZATIONS
10                          CORE
11                          CORE
12                   THUNDERBIRD
13  OTHER_APPLICATIONS_GRAVEYARD
14          FIREFOX_OS_GRAVEYARD

                                    short_description  \
0                            Emergency 2.16.10 Release
1   DOM View is really inefficient with setting wh…
2    Add support for custom cookies and cache headers
3             Create Major Update from 3.0.15 to 3.5.5
4    DCC functionality in ChatZilla isn't functional.
5                                     Fix  and  cruft
6              add a waffle to disable the theme queue
7                               ship ga-IE for 1.5.0.5
8   sidebarheader on GNOME has a transparent backg…
9                         Firefox 3 mk release tracker
10        Bug 200433 regressed by patch in bug 253597
11  View-source highlighting is incorrect for XML …
12  opening a window with -chrome does not work an…
13              Unsightly border around clicked tabs
14  test_user_agent_updates.htmltest_user_agent_ov…

                                    long_description        assignee_name  \
0   2.16.9 is broken -- many users can't enter bug…               mkanat
1   From comment in url:\n\nCurrent code:\nmenuite…              sdwilsh
2   Adding support for custom headers and cookie n…             morgamic
3                                                NaN                catlee
4   User-Agent:       Mozilla/5.0 (Macintosh U PPC…  gijskruitbosch+bugs
5   Since we removed the breadcrumbs and title-ove…                nobody
6   This is a pretty private URL so literally putt…                  kngo
7   the ga-IE locale for Thunderbird 1.5 has owner…               rhelmer
```

8	Created attachment 306847\npatch v1: fixes pro…	myk
9	This is a tracker bug for releasing Firefox 3 …	nobody
10	The patch in bug 253597 regressed the fix from…	mvl
11	User-Agent: Mozilla/5.0 (Windows U Windo…	mrbkap
12	Hi\nI have an extension whose main window is c…	arno
13	With the new skin work (thanks for that Zak!) …	bhsieh
14	See either https://tbpl.mozilla.org/php/getPar…	jimnchen+bmo

	reporter_name	resolution_category	resolution_code	status_category	\
0	mkanat	fixed	1	1	
1	sdwilsh	fixed	1	1	
2	morgamic	fixed	1	1	
3	catlee	fixed	1	1	
4	dafydd	fixed	1	1	
5	jorendorff	fixed	1	1	
6	wclouser	fixed	1	1	
7	mscott	fixed	1	1	
8	myk	fixed	1	1	
9	l10n	fixed	1	1	
10	bzbarsky	fixed	1	1	
11	gilles.ollivier	fixed	1	1	
12	arno	fixed	1	1	
13	zach	fixed	1	1	
14	philringnalda	fixed	1	1	

	status_code	update_date	quantity_of_votes	quantity_of_comments	\
0	4	2005-05-19	0	15	
1	4	2011-06-01	0	8	
2	4	2009-11-02	0	23	
3	4	2013-08-12	0	7	
4	4	2006-02-10	0	14	
5	4	2012-09-18	0	4	
6	4	2014-03-26	0	2	
7	4	2006-07-19	0	12	
8	4	2008-03-12	0	6	
9	4	2008-07-15	0	2	
10	4	2004-09-01	0	8	
11	4	2004-09-17	0	13	
12	4	2009-06-04	0	16	
13	4	2018-10-15	0	4	
14	4	2013-12-06	0	11	

	resolution_date	bug_fix_time	severity_category	severity_code
0	2005-05-19	1	1	6
1	2007-01-14	36	0	2
2	2008-03-24	174	1	6
3	2009-11-10	8	0	2

```
4       2005-12-07             5                  0              2
5       2008-02-12             1                  0              2
6       2014-03-13            14                  0              2
7       2006-06-26            26                  0              2
8       2008-03-05             3                  0              2
9       2008-07-15           162                  0              2
10      2004-09-01            27                  4              4
11      2004-09-17           501                  0              2
12      2008-01-15             0                  0              2
13      2007-08-14             6                  0              2
14      2013-12-06            13                  0              2
```

[ ]:
```
mbrd['component_name'].unique
```

[ ]:
```
<bound method Series.unique of 0       Bugzilla-General
1          DOM Inspector
2                General
3                General
4              ChatZilla
               …
9994             General
9995               XPCOM
9996             General
9997     fy-NL / Frisian
9998    Spelling checker
Name: component_name, Length: 9999, dtype: object>
```

[ ]:
```
mbrd['product_name'].unique
```

[ ]:
```
<bound method Series.unique of 0                    BUGZILLA
1          OTHER_APPLICATIONS
2         SUPPORT.MOZILLA.ORG
3         RELEASE_ENGINEERING
4          OTHER_APPLICATIONS
                …
9994          WWW.MOZILLA.ORG
9995                     CORE
9996     FIREFOX_BUILD_SYSTEM
9997    MOZILLA_LOCALIZATIONS
9998                     CORE
Name: product_name, Length: 9999, dtype: object>
```

1.6 As other elements are not having a categorical

1.6.1 Now we will be removing null values

[ ]:
```
mbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       1
long_description       79
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
# percentage of missing values
mbrd.isnull().sum() / mbrd.shape[0] * 100
```

[ ]:
```
bug_id                  0.000000
creation_date           0.000000
component_name          0.000000
product_name            0.000000
short_description       0.010001
long_description        0.790079
assignee_name           0.000000
reporter_name           0.000000
resolution_category     0.000000
resolution_code         0.000000
status_category         0.000000
status_code             0.000000
update_date             0.000000
quantity_of_votes       0.000000
quantity_of_comments    0.000000
resolution_date         0.000000
bug_fix_time            0.000000
severity_category       0.000000
```

```
severity_code           0.000000
dtype: float64
```

```
[ ]: mbrd.dropna(inplace=True)
```

```
[ ]: mbrd.isnull().sum()
```

```
[ ]: bug_id                 0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description        0
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

```
[ ]: mbrd.head()
```

```
[ ]:                        bug_id creation_date    component_name  \
0              BUGZILLA-294734    2005-05-18  Bugzilla-General
1    OTHER_APPLICATIONS-363323    2006-12-09     DOM Inspector
2   SUPPORT.MOZILLA.ORG-398246    2007-10-02           General
4    OTHER_APPLICATIONS-318859    2005-12-02         ChatZilla
5  DEVELOPER.MOZILLA.ORG-416840    2008-02-11           General

            product_name                                short_description  \
0              BUGZILLA                          Emergency 2.16.10 Release
1    OTHER_APPLICATIONS  DOM View is really inefficient with setting wh…
2   SUPPORT.MOZILLA.ORG   Add support for custom cookies and cache headers
4    OTHER_APPLICATIONS   DCC functionality in ChatZilla isn't functional.
5  DEVELOPER.MOZILLA.ORG                                  Fix  and  cruft

                                long_description        assignee_name  \
0  2.16.9 is broken -- many users can't enter bug…              mkanat
1  From comment in url:\n\nCurrent code:\nmenuite…             sdwilsh
```

```
2  Adding support for custom headers and cookie n…             morgamic
4  User-Agent:       Mozilla/5.0 (Macintosh U PPC…  gijskruitbosch+bugs
5  Since we removed the breadcrumbs and title-ove…               nobody
```

	reporter_name	resolution_category	resolution_code	status_category	status_code	update_date	quantity_of_votes	quantity_of_comments	resolution_date	bug_fix_time	severity_category	severity_code
0	mkanat	fixed	1	1	4	2005-05-19	0	15	2005-05-19	1	1	6
1	sdwilsh	fixed	1	1	4	2011-06-01	0	8	2007-01-14	36	0	2
2	morgamic	fixed	1	1	4	2009-11-02	0	23	2008-03-24	174	1	6
4	dafydd	fixed	1	1	4	2006-02-10	0	14	2005-12-07	5	0	2
5	jorendorff	fixed	1	1	4	2012-09-18	0	4	2008-02-12	1	0	2

[ ]:
```
from datetime import time,date,datetime

mbrd['creation_date'] = pd.to_datetime(mbrd['creation_date'])
mbrd['update_date'] = pd.to_datetime(mbrd['update_date'])
mbrd['resolution_date'] = pd.to_datetime(mbrd['resolution_date'])
```

[ ]:
```
mbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9919 entries, 0 to 9998
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   bug_id               9919 non-null   object
 1   creation_date        9919 non-null   datetime64[ns]
 2   component_name       9919 non-null   object
 3   product_name         9919 non-null   object
 4   short_description    9919 non-null   object
 5   long_description     9919 non-null   object
 6   assignee_name        9919 non-null   object
 7   reporter_name        9919 non-null   object
 8   resolution_category  9919 non-null   object
```

```
 9   resolution_code       9919 non-null   int64
 10  status_category       9919 non-null   int64
 11  status_code           9919 non-null   int64
 12  update_date           9919 non-null   datetime64[ns]
 13  quantity_of_votes     9919 non-null   int64
 14  quantity_of_comments  9919 non-null   int64
 15  resolution_date       9919 non-null   datetime64[ns]
 16  bug_fix_time          9919 non-null   int64
 17  severity_category     9919 non-null   int64
 18  severity_code         9919 non-null   int64
dtypes: datetime64[ns](3), int64(8), object(8)
memory usage: 1.5+ MB
```

[ ]:
```
mbrd.insert(0, "time_to_dep[s]", ((
    mbrd['resolution_date']-mbrd['creation_date']).astype('timedelta64[s]')),␣
 ↪True)
```

[ ]:
```
mbrd.head()
```

[ ]:
```
   time_to_dep[s]                      bug_id creation_date  \
0         86400.0               BUGZILLA-294734    2005-05-18
1       3110400.0    OTHER_APPLICATIONS-363323    2006-12-09
2      15033600.0   SUPPORT.MOZILLA.ORG-398246    2007-10-02
4        432000.0    OTHER_APPLICATIONS-318859    2005-12-02
5         86400.0 DEVELOPER.MOZILLA.ORG-416840    2008-02-11

     component_name          product_name  \
0  Bugzilla-General               BUGZILLA
1     DOM Inspector    OTHER_APPLICATIONS
2           General   SUPPORT.MOZILLA.ORG
4         ChatZilla    OTHER_APPLICATIONS
5           General  DEVELOPER.MOZILLA.ORG

                                short_description  \
0                          Emergency 2.16.10 Release
1  DOM View is really inefficient with setting wh…
2   Add support for custom cookies and cache headers
4   DCC functionality in ChatZilla isn't functional.
5                                  Fix  and  cruft

                                long_description       assignee_name  \
0  2.16.9 is broken -- many users can't enter bug…              mkanat
1  From comment in url:\n\nCurrent code:\nmenuite…             sdwilsh
2  Adding support for custom headers and cookie n…            morgamic
4  User-Agent:       Mozilla/5.0 (Macintosh U PPC…  gijskruitbosch+bugs
5  Since we removed the breadcrumbs and title-ove…              nobody
```

```
  reporter_name resolution_category  resolution_code  status_category  \
0        mkanat               fixed               1                1
1       sdwilsh               fixed               1                1
2      morgamic               fixed               1                1
4        dafydd               fixed               1                1
5    jorendorff               fixed               1                1

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            4  2005-05-19                  0                    15
1            4  2011-06-01                  0                     8
2            4  2009-11-02                  0                    23
4            4  2006-02-10                  0                    14
5            4  2012-09-18                  0                     4

  resolution_date  bug_fix_time  severity_category  severity_code
0      2005-05-19             1                  1              6
1      2007-01-14            36                  0              2
2      2008-03-24           174                  1              6
4      2005-12-07             5                  0              2
5      2008-02-12             1                  0              2
```

[ ]:
```
# dropping extra columns
mbrd.
 ↪drop(['bug_id','resolution_code','creation_date','component_name','product_name','short_des
```

[ ]:
```
# now we are done with dropping values as well
mbrd['bug_fix_time'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew         4.065429
kurtosis    20.980392
Name: bug_fix_time, dtype: float64
```

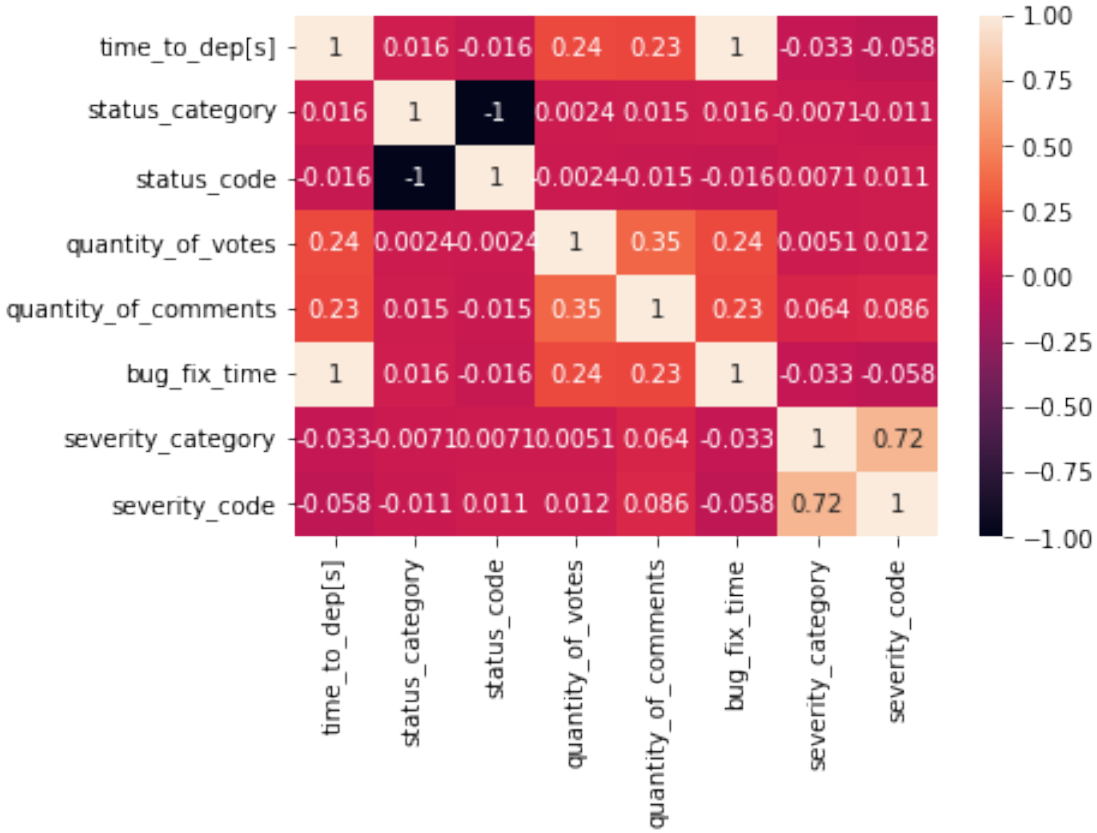
[ ]:
```
corr = mbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.016368    -0.016368
status_category             0.016368         1.000000    -1.000000
status_code                -0.016368        -1.000000     1.000000
quantity_of_votes           0.240403         0.002442    -0.002442
quantity_of_comments        0.230322         0.014653    -0.014653
bug_fix_time                1.000000         0.016368    -0.016368
severity_category          -0.033380        -0.007136     0.007136
severity_code              -0.057978        -0.010556     0.010556

                      quantity_of_votes  quantity_of_comments  bug_fix_time  \
time_to_dep[s]                 0.240403              0.230322      1.000000
```

```
status_category              0.002442              0.014653      0.016368
status_code                 -0.002442             -0.014653     -0.016368
quantity_of_votes            1.000000              0.354797      0.240403
quantity_of_comments         0.354797              1.000000      0.230322
bug_fix_time                 0.240403              0.230322      1.000000
severity_category            0.005063              0.063767     -0.033380
severity_code                0.011688              0.085559     -0.057978

                     severity_category  severity_code
time_to_dep[s]               -0.033380      -0.057978
status_category              -0.007136      -0.010556
status_code                   0.007136       0.010556
quantity_of_votes             0.005063       0.011688
quantity_of_comments          0.063767       0.085559
bug_fix_time                 -0.033380      -0.057978
severity_category             1.000000       0.722919
severity_code                 0.722919       1.000000
```

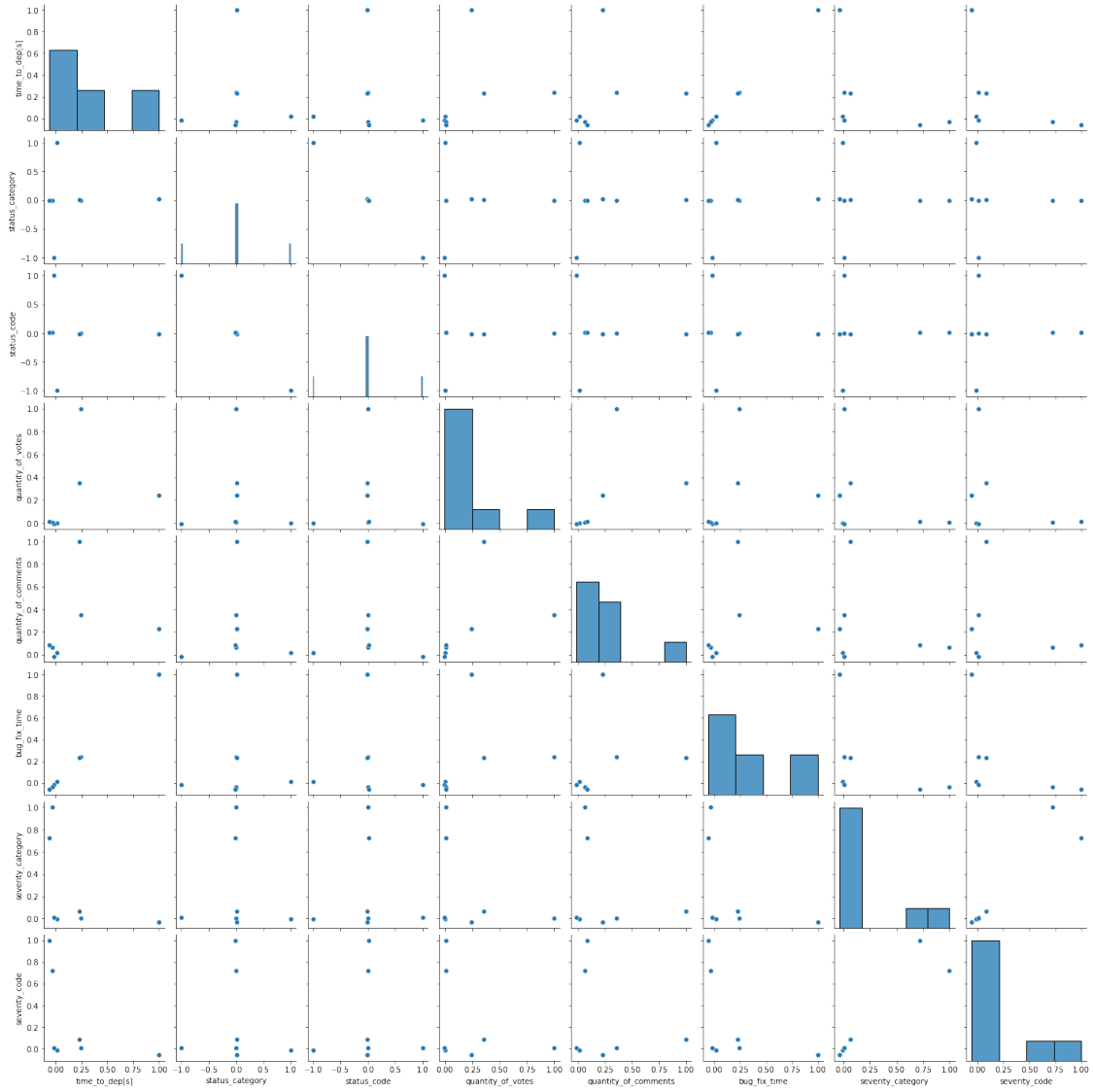
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]: <AxesSubplot:>



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]: <seaborn.axisgrid.PairGrid at 0x266b3681a20>



[ ]:
```
mbrd.head()
```

[ ]:

	time_to_dep[s]	status_category	status_code	quantity_of_votes	\
0	86400.0	1	4	0	
1	3110400.0	1	4	0	
2	15033600.0	1	4	0	

```
4        432000.0                1              4                   0
5         86400.0                1              4                   0

   quantity_of_comments  bug_fix_time  severity_category  severity_code
0                    15             1                  1              6
1                     8            36                  0              2
2                    23           174                  1              6
4                    14             5                  0              2
5                     4             1                  0              2
```

[ ]:
```
X = mbrd.iloc[:,:-1].values #rows and then columns in brackets
Y = mbrd.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[8.64000e+04, 1.00000e+00, 4.00000e+00, …, 1.50000e+01,
        1.00000e+00, 1.00000e+00],
       [3.11040e+06, 1.00000e+00, 4.00000e+00, …, 8.00000e+00,
        3.60000e+01, 0.00000e+00],
       [1.50336e+07, 1.00000e+00, 4.00000e+00, …, 2.30000e+01,
        1.74000e+02, 1.00000e+00],
       …,
       [8.64000e+05, 1.00000e+00, 4.00000e+00, …, 1.50000e+01,
        1.00000e+01, 0.00000e+00],
       [6.50592e+07, 1.00000e+00, 4.00000e+00, …, 3.00000e+00,
        7.53000e+02, 0.00000e+00],
       [3.02400e+06, 1.00000e+00, 4.00000e+00, …, 6.00000e+00,
        3.50000e+01, 3.00000e+00]])
```

[ ]:
```
Y
```

[ ]:
```
array([6, 2, 6, …, 2, 2, 2], dtype=int64)
```

1.6.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
#
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=42)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
        # assignment which one we should accept from these
```

```
LinearRegression()
R2 score is:  0.4920245419541416
Mean Absolute error is:  0.31309511788320527
Mean Squared Error:  0.5857018758661755
---------------------------------------------------
LogisticRegression()
R2 score is:  -0.15362281926544386
Mean Absolute error is:  0.47731854838709675
Mean Squared Error:  1.330141129032258
---------------------------------------------------
GaussianNB()
R2 score is:  -0.15362281926544386
Mean Absolute error is:  0.47731854838709675
Mean Squared Error:  1.330141129032258
---------------------------------------------------
RandomForestClassifier()
R2 score is:  0.9890714018637226
Mean Absolute error is:  0.0025201612903225806
Mean Squared Error:  0.012600806451612902
---------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
```

```
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
--------------------------------------------------
SVR()
R2 score is:  -0.08927021753849762
Mean Absolute error is:  0.544030279351995
Mean Squared Error:  1.2559417972508817
--------------------------------------------------
```

1.6.3 EDA On Sales_data_sample

[ ]:
```
sds.head()
```

[ ]:
```
   ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
0        10107               30      95.70                2  2871.00
1        10121               34      81.35                5  2765.90
2        10134               41      94.74                2  3884.34
3        10145               45      83.26                6  3746.70
4        10159               49     100.00               14  5205.27

          ORDERDATE   STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  \
0   2/24/2003 0:00  Shipped       1         2     2003  …
1    5/7/2003 0:00  Shipped       2         5     2003  …
2    7/1/2003 0:00  Shipped       3         7     2003  …
3   8/25/2003 0:00  Shipped       3         8     2003  …
4  10/10/2003 0:00  Shipped       4        10     2003  …

                     ADDRESSLINE1  ADDRESSLINE2           CITY STATE  \
0         897 Long Airport Avenue           NaN            NYC    NY
1              59 rue de l'Abbaye           NaN          Reims   NaN
2  27 rue du Colonel Pierre Avia           NaN          Paris   NaN
3              78934 Hillside Dr.           NaN       Pasadena    CA
4                 7734 Strong St.           NaN  San Francisco    CA

  POSTALCODE COUNTRY TERRITORY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
0      10022     USA       NaN             Yu             Kwai    Small
1      51100  France      EMEA        Henriot             Paul    Small
2      75508  France      EMEA       Da Cunha           Daniel   Medium
3      90003     USA       NaN          Young            Julie   Medium
4        NaN     USA       NaN          Brown            Julie   Medium

[5 rows x 25 columns]
```

[ ]:
```
sds.tail()
```

[ ]:
```
      ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
2818        10350               20     100.00               15  2244.40
```

```
2819        10373              29      100.00               1  3978.51
2820        10386              43      100.00               4  5417.57
2821        10397              34       62.24               1  2116.16
2822        10414              47       65.52               9  3079.44

          ORDERDATE    STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  \
2818  12/2/2004 0:00   Shipped       4        12     2004  …
2819  1/31/2005 0:00   Shipped       1         1     2005  …
2820   3/1/2005 0:00  Resolved       1         3     2005  …
2821  3/28/2005 0:00   Shipped       1         3     2005  …
2822   5/6/2005 0:00   On Hold       2         5     2005  …

              ADDRESSLINE1  ADDRESSLINE2      CITY STATE POSTALCODE  COUNTRY  \
2818     C/ Moralzarzal, 86           NaN    Madrid   NaN      28034    Spain
2819           Torikatu 38           NaN      Oulu   NaN      90110  Finland
2820     C/ Moralzarzal, 86           NaN    Madrid   NaN      28034    Spain
2821  1 rue Alsace-Lorraine           NaN  Toulouse   NaN      31000   France
2822     8616 Spinnaker Dr.           NaN    Boston    MA      51003      USA

     TERRITORY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
2818      EMEA          Freyre            Diego    Small
2819      EMEA       Koskitalo           Pirkko   Medium
2820      EMEA          Freyre            Diego   Medium
2821      EMEA          Roulet          Annette    Small
2822       NaN         Yoshido             Juri   Medium

[5 rows x 25 columns]
```

[ ]:
```
sds.describe()
```

[ ]:
```
        ORDERNUMBER  QUANTITYORDERED    PRICEEACH  ORDERLINENUMBER  \
count   2823.000000      2823.000000  2823.000000      2823.000000
mean   10258.725115        35.092809    83.658544         6.466171
std       92.085478         9.741443    20.174277         4.225841
min    10100.000000         6.000000    26.880000         1.000000
25%    10180.000000        27.000000    68.860000         3.000000
50%    10262.000000        35.000000    95.700000         6.000000
75%    10333.500000        43.000000   100.000000         9.000000
max    10425.000000        97.000000   100.000000        18.000000

              SALES       QTR_ID     MONTH_ID     YEAR_ID         MSRP
count   2823.000000  2823.000000  2823.000000  2823.00000  2823.000000
mean    3553.889072     2.717676     7.092455  2003.81509   100.715551
std     1841.865106     1.203878     3.656633     0.69967    40.187912
min      482.130000     1.000000     1.000000  2003.00000    33.000000
25%     2203.430000     2.000000     4.000000  2003.00000    68.000000
50%     3184.800000     3.000000     8.000000  2004.00000    99.000000
```

```
75%     4508.000000      4.000000     11.000000  2004.00000   124.000000
max    14082.800000      4.000000     12.000000  2005.00000   214.000000
```

[ ]:
```
sds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
Data columns (total 25 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ORDERNUMBER       2823 non-null   int64
 1   QUANTITYORDERED   2823 non-null   int64
 2   PRICEEACH         2823 non-null   float64
 3   ORDERLINENUMBER   2823 non-null   int64
 4   SALES             2823 non-null   float64
 5   ORDERDATE         2823 non-null   object
 6   STATUS            2823 non-null   object
 7   QTR_ID            2823 non-null   int64
 8   MONTH_ID          2823 non-null   int64
 9   YEAR_ID           2823 non-null   int64
 10  PRODUCTLINE       2823 non-null   object
 11  MSRP              2823 non-null   int64
 12  PRODUCTCODE       2823 non-null   object
 13  CUSTOMERNAME      2823 non-null   object
 14  PHONE             2823 non-null   object
 15  ADDRESSLINE1      2823 non-null   object
 16  ADDRESSLINE2      302 non-null    object
 17  CITY              2823 non-null   object
 18  STATE             1337 non-null   object
 19  POSTALCODE        2747 non-null   object
 20  COUNTRY           2823 non-null   object
 21  TERRITORY         1749 non-null   object
 22  CONTACTLASTNAME   2823 non-null   object
 23  CONTACTFIRSTNAME  2823 non-null   object
 24  DEALSIZE          2823 non-null   object
dtypes: float64(2), int64(7), object(16)
memory usage: 551.5+ KB
```

[ ]:
```
sds['STATUS'].unique()
```

[ ]:
```
array(['Shipped', 'Disputed', 'In Process', 'Cancelled', 'On Hold',
       'Resolved'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status we can encode it as it is two categories only
# We are encoding it
sds['STATUS'] = sds['STATUS'].replace("Cancelled", 0)
```

```
sds['STATUS'] = sds['STATUS'].replace("Shipped", 1)
sds['STATUS'] = sds['STATUS'].replace("Disputed", 2)
sds['STATUS'] = sds['STATUS'].replace("In Process", 3)
sds['STATUS'] = sds['STATUS'].replace("On Hold", 4)
sds['STATUS'] = sds['STATUS'].replace("Resolved", 5)
```

[ ]:
```
sds['TERRITORY'].unique()
```

[ ]:
```
array([nan, 'EMEA', 'APAC', 'Japan'], dtype=object)
```

[ ]:
```
# As we have another column TERRITORY and we can perform encoding on it as well
sds['TERRITORY'] = sds['TERRITORY'].replace("NaN", 0)
sds['TERRITORY'] = sds['TERRITORY'].replace("EMEA", 1)
sds['TERRITORY'] = sds['TERRITORY'].replace("APAC", 2)
sds['TERRITORY'] = sds['TERRITORY'].replace("Japan", 3)
```

[ ]:
```
sds['PRODUCTLINE'].unique()
```

[ ]:
```
array(['Motorcycles', 'Classic Cars', 'Trucks and Buses', 'Vintage Cars',
       'Planes', 'Ships', 'Trains'], dtype=object)
```

[ ]:
```
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Motorcycles", 0)
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Classic Cars", 1)
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Trucks and Buses", 2)
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Vintage Cars", 3)
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Planes", 4)
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Ships", 5)
sds['PRODUCTLINE'] = sds['PRODUCTLINE'].replace("Trains", 6)
```

[ ]:
```
sds.head(15)
```

[ ]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	\
0	10107	30	95.70	2	2871.00	
1	10121	34	81.35	5	2765.90	
2	10134	41	94.74	2	3884.34	
3	10145	45	83.26	6	3746.70	
4	10159	49	100.00	14	5205.27	
5	10168	36	96.66	1	3479.76	
6	10180	29	86.13	9	2497.77	
7	10188	48	100.00	1	5512.32	
8	10201	22	98.57	2	2168.54	
9	10211	41	100.00	14	4708.44	
10	10223	37	100.00	1	3965.66	
11	10237	23	100.00	7	2333.12	
12	10251	28	100.00	2	3188.64	
13	10263	34	100.00	2	3676.76	
14	10275	45	92.83	1	4177.35	

```
          ORDERDATE  STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  \
0    2/24/2003 0:00       1       1         2     2003  …
1     5/7/2003 0:00       1       2         5     2003  …
2     7/1/2003 0:00       1       3         7     2003  …
3    8/25/2003 0:00       1       3         8     2003  …
4   10/10/2003 0:00       1       4        10     2003  …
5   10/28/2003 0:00       1       4        10     2003  …
6   11/11/2003 0:00       1       4        11     2003  …
7   11/18/2003 0:00       1       4        11     2003  …
8    12/1/2003 0:00       1       4        12     2003  …
9    1/15/2004 0:00       1       1         1     2004  …
10   2/20/2004 0:00       1       1         2     2004  …
11    4/5/2004 0:00       1       2         4     2004  …
12   5/18/2004 0:00       1       2         5     2004  …
13   6/28/2004 0:00       1       2         6     2004  …
14   7/23/2004 0:00       1       3         7     2004  …

                    ADDRESSLINE1  ADDRESSLINE2           CITY     STATE  \
0          897 Long Airport Avenue           NaN            NYC        NY
1               59 rue de l'Abbaye           NaN          Reims       NaN
2   27 rue du Colonel Pierre Avia           NaN          Paris       NaN
3              78934 Hillside Dr.           NaN       Pasadena        CA
4                 7734 Strong St.           NaN  San Francisco        CA
5               9408 Furth Circle           NaN     Burlingame        CA
6          184, chausse de Tournai           NaN          Lille       NaN
7     Drammen 121, PR 744 Sentrum            NaN         Bergen       NaN
8       5557 North Pendale Street           NaN  San Francisco        CA
9                25, rue Lauriston           NaN          Paris       NaN
10               636 St Kilda Road      Level 3      Melbourne  Victoria
11               2678 Kingston Rd.     Suite 101            NYC        NY
12                   7476 Moss Rd.           NaN         Newark        NJ
13             25593 South Bay Ln.           NaN    Bridgewater        CT
14   67, rue des Cinquante Otages            NaN         Nantes       NaN

   POSTALCODE    COUNTRY TERRITORY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
0       10022        USA       NaN              Yu             Kwai    Small
1       51100     France       1.0         Henriot             Paul    Small
2       75508     France       1.0        Da Cunha           Daniel   Medium
3       90003        USA       NaN           Young            Julie   Medium
4         NaN        USA       NaN           Brown            Julie   Medium
5       94217        USA       NaN          Hirano             Juri   Medium
6       59000     France       1.0           Rance          Martine    Small
7      N 5804     Norway       1.0          Oeztan           Veysel   Medium
8         NaN        USA       NaN          Murphy            Julie    Small
9       75016     France       1.0         Perrier        Dominique   Medium
10       3004  Australia       2.0        Ferguson            Peter   Medium
```

```
11      10022        USA       NaN           Frick          Michael    Small
12      94019        USA       NaN           Brown          William   Medium
13      97562        USA       NaN            King            Julie   Medium
14      44000     France       1.0         Labrune           Janine   Medium

[15 rows x 25 columns]
```

```
[ ]: sds['DEALSIZE'].unique()
```

```
[ ]: array(['Small', 'Medium', 'Large'], dtype=object)
```

```
[ ]: sds['DEALSIZE'] = sds['DEALSIZE'].replace("Small", 0)
     sds['DEALSIZE'] = sds['DEALSIZE'].replace("Medium", 1)
     sds['DEALSIZE'] = sds['DEALSIZE'].replace("Large", 2)
```

```
[ ]: sds['CONTACTLASTNAME'].unique()
```

```
[ ]: array(['Yu', 'Henriot', 'Da Cunha', 'Young', 'Brown', 'Hirano', 'Rance',
            'Oeztan', 'Murphy', 'Perrier', 'Ferguson', 'Frick', 'King',
            'Labrune', 'Hernandez', 'Karttunen', 'Bergulfsen', 'Pipps',
            'Huxley', 'Benitez', 'Devon', 'Freyre', 'Berglund', 'Sommer',
            'Natividad', 'Calaghan', 'Cervantes', 'Saveley', 'Tannamuri',
            'Thompson', 'Tseng', 'Shimamura', 'Accorti', 'Larsson', 'Tonini',
            'Nelson', "O'Hara", 'Fresnisre', 'Kentary', 'Schmitt', 'Petersen',
            'Tam', 'Roulet', 'Hardy', 'Saavedra', 'Chandler', 'Dewey',
            'Lincoln', 'Yoshido', 'Bennett', 'Koskitalo', 'Bertrand', 'Mendel',
            'Franco', 'Victorino', 'Cruz', 'Lebihan', 'Pfalzheim', 'Holz',
            'Moroni', 'Barajas', 'Keitel', 'Suominen', 'Cassidy', 'Fernandez',
            'Ashworth', 'Ibsen', 'Kuo', 'Roel', 'Taylor', 'Citeaux', 'Klaeboe',
            'Rovelli', 'Connery', 'Lewis', 'Donnermeyer', 'Cartrain'],
           dtype=object)
```

1.7 As other elements are not having a categorical

1.7.1 Now we will be removing null values

```
[ ]: sds.isnull().sum()
```

```
[ ]: ORDERNUMBER          0
     QUANTITYORDERED     0
     PRICEEACH           0
     ORDERLINENUMBER     0
     SALES               0
     ORDERDATE           0
     STATUS              0
     QTR_ID              0
     MONTH_ID            0
```

```
YEAR_ID               0
PRODUCTLINE           0
MSRP                  0
PRODUCTCODE           0
CUSTOMERNAME          0
PHONE                 0
ADDRESSLINE1          0
ADDRESSLINE2       2521
CITY                  0
STATE              1486
POSTALCODE           76
COUNTRY               0
TERRITORY          1074
CONTACTLASTNAME       0
CONTACTFIRSTNAME      0
DEALSIZE              0
dtype: int64
```

[ ]:
```
# percentage of missing values
sds.isnull().sum() / sds.shape[0] * 100
```

[ ]:
```
ORDERNUMBER         0.000000
QUANTITYORDERED     0.000000
PRICEEACH           0.000000
ORDERLINENUMBER     0.000000
SALES               0.000000
ORDERDATE           0.000000
STATUS              0.000000
QTR_ID              0.000000
MONTH_ID            0.000000
YEAR_ID             0.000000
PRODUCTLINE         0.000000
MSRP                0.000000
PRODUCTCODE         0.000000
CUSTOMERNAME        0.000000
PHONE               0.000000
ADDRESSLINE1        0.000000
ADDRESSLINE2       89.302161
CITY                0.000000
STATE              52.639036
POSTALCODE          2.692171
COUNTRY             0.000000
TERRITORY          38.044633
CONTACTLASTNAME     0.000000
CONTACTFIRSTNAME    0.000000
DEALSIZE            0.000000
dtype: float64
```

```
[ ]: # now we will be removes those rows which have most null values and are not in␣
      ↪our use
     sds.drop("ADDRESSLINE2", axis=1, inplace=True)
```

```
[ ]: sds.drop("STATE", axis=1, inplace=True)
```

```
[ ]: sds.drop("TERRITORY", axis=1, inplace=True)
```

```
[ ]: sds.dropna(inplace=True)
```

```
[ ]: sds.isnull().sum()
```

```
[ ]: ORDERNUMBER         0
     QUANTITYORDERED     0
     PRICEEACH           0
     ORDERLINENUMBER     0
     SALES               0
     ORDERDATE           0
     STATUS              0
     QTR_ID              0
     MONTH_ID            0
     YEAR_ID             0
     PRODUCTLINE         0
     MSRP                0
     PRODUCTCODE         0
     CUSTOMERNAME        0
     PHONE               0
     ADDRESSLINE1        0
     CITY                0
     POSTALCODE          0
     COUNTRY             0
     CONTACTLASTNAME     0
     CONTACTFIRSTNAME    0
     DEALSIZE            0
     dtype: int64
```

```
[ ]: sds.head()
```

```
[ ]:    ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
     0        10107                30      95.70                2  2871.00
     1        10121                34      81.35                5  2765.90
     2        10134                41      94.74                2  3884.34
     3        10145                45      83.26                6  3746.70
     5        10168                36      96.66                1  3479.76

              ORDERDATE  STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  PRODUCTCODE  \
     0   2/24/2003 0:00       1       1         2     2003  …     S10_1678
```

```
1    5/7/2003 0:00       1       2         5     2003  …     S10_1678
2    7/1/2003 0:00       1       3         7     2003  …     S10_1678
3   8/25/2003 0:00       1       3         8     2003  …     S10_1678
5  10/28/2003 0:00       1       4        10     2003  …     S10_1678

           CUSTOMERNAME            PHONE                  ADDRESSLINE1  \
0     Land of Toys Inc.        2125557818        897 Long Airport Avenue
1    Reims Collectables        26.47.1555             59 rue de l'Abbaye
2       Lyon Souveniers  +33 1 46 62 7555  27 rue du Colonel Pierre Avia
3     Toys4GrownUps.com        6265557265             78934 Hillside Dr.
5  Technics Stores Inc.        6505556809              9408 Furth Circle

          CITY POSTALCODE COUNTRY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
0          NYC      10022     USA              Yu             Kwai        0
1       Reims      51100  France         Henriot             Paul        0
2       Paris      75508  France        Da Cunha           Daniel        1
3    Pasadena      90003     USA           Young            Julie        1
5  Burlingame      94217     USA          Hirano             Juri        1

[5 rows x 22 columns]
```

[ ]:
```
from datetime import time,date,datetime

sds['ORDERDATE'] = pd.to_datetime(sds['ORDERDATE'])
```

[ ]:
```
sds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2747 entries, 0 to 2822
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ORDERNUMBER       2747 non-null   int64
 1   QUANTITYORDERED   2747 non-null   int64
 2   PRICEEACH         2747 non-null   float64
 3   ORDERLINENUMBER   2747 non-null   int64
 4   SALES             2747 non-null   float64
 5   ORDERDATE         2747 non-null   datetime64[ns]
 6   STATUS            2747 non-null   int64
 7   QTR_ID            2747 non-null   int64
 8   MONTH_ID          2747 non-null   int64
 9   YEAR_ID           2747 non-null   int64
 10  PRODUCTLINE       2747 non-null   int64
 11  MSRP              2747 non-null   int64
 12  PRODUCTCODE       2747 non-null   object
 13  CUSTOMERNAME      2747 non-null   object
 14  PHONE             2747 non-null   object
```

```
 15  ADDRESSLINE1      2747 non-null   object
 16  CITY              2747 non-null   object
 17  POSTALCODE        2747 non-null   object
 18  COUNTRY           2747 non-null   object
 19  CONTACTLASTNAME   2747 non-null   object
 20  CONTACTFIRSTNAME  2747 non-null   object
 21  DEALSIZE          2747 non-null   int64
dtypes: datetime64[ns](1), float64(2), int64(10), object(9)
memory usage: 493.6+ KB
```

[ ]:
```
sds.head()
```

[ ]:
```
   ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER    SALES  \
0        10107               30      95.70                2  2871.00
1        10121               34      81.35                5  2765.90
2        10134               41      94.74                2  3884.34
3        10145               45      83.26                6  3746.70
5        10168               36      96.66                1  3479.76

   ORDERDATE  STATUS  QTR_ID  MONTH_ID  YEAR_ID  …  PRODUCTCODE  \
0 2003-02-24       1       1         2     2003  …     S10_1678
1 2003-05-07       1       2         5     2003  …     S10_1678
2 2003-07-01       1       3         7     2003  …     S10_1678
3 2003-08-25       1       3         8     2003  …     S10_1678
5 2003-10-28       1       4        10     2003  …     S10_1678

           CUSTOMERNAME             PHONE                  ADDRESSLINE1  \
0     Land of Toys Inc.        2125557818        897 Long Airport Avenue
1    Reims Collectables         26.47.1555             59 rue de l'Abbaye
2       Lyon Souveniers  +33 1 46 62 7555  27 rue du Colonel Pierre Avia
3     Toys4GrownUps.com        6265557265             78934 Hillside Dr.
5  Technics Stores Inc.        6505556809              9408 Furth Circle

         CITY POSTALCODE COUNTRY CONTACTLASTNAME CONTACTFIRSTNAME DEALSIZE
0         NYC      10022     USA              Yu              Kwai        0
1       Reims      51100  France         Henriot              Paul        0
2       Paris      75508  France        Da Cunha            Daniel        1
3    Pasadena      90003     USA           Young             Julie        1
5  Burlingame      94217     USA          Hirano              Juri        1

[5 rows x 22 columns]
```

[ ]:
```
# dropping extra columns
sds.drop(['PRODUCTCODE', 'CUSTOMERNAME', 'PHONE', 'ADDRESSLINE1', 'CITY',␣
 ↪'COUNTRY',
        'CONTACTLASTNAME', 'CONTACTFIRSTNAME', 'ORDERDATE', 'POSTALCODE'],␣
 ↪axis=1, inplace=True)
```

[ ]:
```
# now we are done with dropping values as well
sds['ORDERNUMBER'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew       -0.006995
kurtosis   -1.154407
Name: ORDERNUMBER, dtype: float64
```

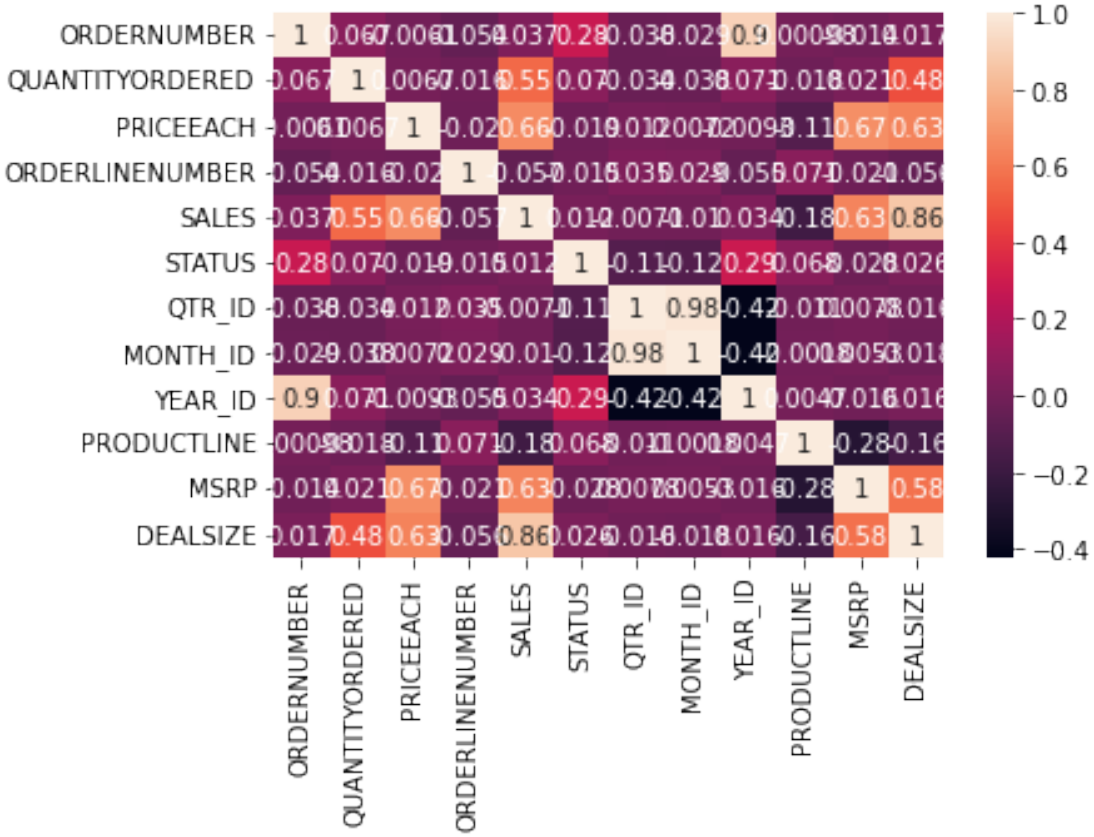
[ ]:
```
corr = sds.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                 ORDERNUMBER  QUANTITYORDERED  PRICEEACH  ORDERLINENUMBER  \
ORDERNUMBER         1.000000         0.067110  -0.006073        -0.054300
QUANTITYORDERED     0.067110         1.000000   0.006671        -0.016295
PRICEEACH          -0.006073         0.006671   1.000000        -0.020478
ORDERLINENUMBER    -0.054300        -0.016295  -0.020478         1.000000
SALES               0.037289         0.553359   0.658012        -0.057414
STATUS              0.279049         0.070457  -0.019400        -0.014889
QTR_ID             -0.037954        -0.034440   0.011677         0.034895
MONTH_ID           -0.028515        -0.037926   0.007213         0.029180
YEAR_ID             0.903582         0.070520  -0.009327        -0.055058
PRODUCTLINE         0.000980        -0.018475  -0.108890         0.071391
MSRP               -0.013910         0.020551   0.669348        -0.020956
DEALSIZE            0.017148         0.477134   0.629671        -0.056093

                      SALES    STATUS    QTR_ID  MONTH_ID   YEAR_ID  \
ORDERNUMBER        0.037289  0.279049 -0.037954 -0.028515  0.903582
QUANTITYORDERED    0.553359  0.070457 -0.034440 -0.037926  0.070520
PRICEEACH          0.658012 -0.019400  0.011677  0.007213 -0.009327
ORDERLINENUMBER   -0.057414 -0.014889  0.034895  0.029180 -0.055058
SALES              1.000000  0.011938 -0.007119 -0.010200  0.033604
STATUS             0.011938  1.000000 -0.106929 -0.117196  0.285451
QTR_ID            -0.007119 -0.106929  1.000000  0.979257 -0.423081
MONTH_ID          -0.010200 -0.117196  0.979257  1.000000 -0.421548
YEAR_ID            0.033604  0.285451 -0.423081 -0.421548  1.000000
PRODUCTLINE       -0.179857  0.068094 -0.010580 -0.001786  0.004652
MSRP               0.634849 -0.028374  0.007792  0.005306 -0.016434
DEALSIZE           0.861707  0.026224 -0.016155 -0.017534  0.015596

                 PRODUCTLINE      MSRP  DEALSIZE
ORDERNUMBER         0.000980 -0.013910  0.017148
QUANTITYORDERED    -0.018475  0.020551  0.477134
PRICEEACH          -0.108890  0.669348  0.629671
ORDERLINENUMBER     0.071391 -0.020956 -0.056093
SALES              -0.179857  0.634849  0.861707
STATUS              0.068094 -0.028374  0.026224
QTR_ID             -0.010580  0.007792 -0.016155
MONTH_ID           -0.001786  0.005306 -0.017534
```

```
YEAR_ID            0.004652 -0.016434  0.015596
PRODUCTLINE        1.000000 -0.278800 -0.162668
MSRP              -0.278800  1.000000  0.576289
DEALSIZE          -0.162668  0.576289  1.000000
```

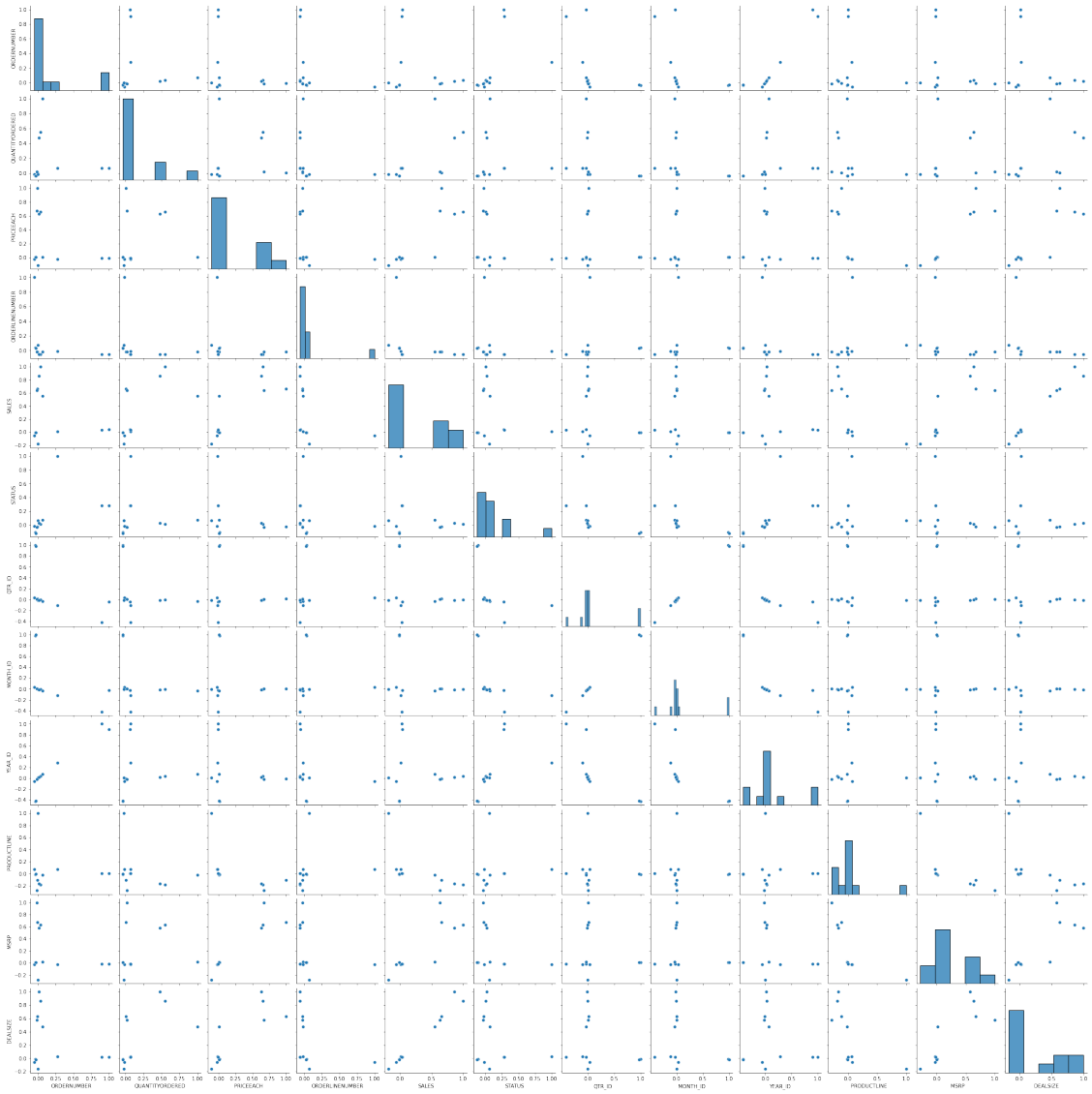
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]:
```
<AxesSubplot:>
```



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]:
```
<seaborn.axisgrid.PairGrid at 0x266b834bcd0>
```



[ ]:
```
sds.head()
```

[ ]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	STATUS	\
0	10107	30	95.70	2	2871.00	1	
1	10121	34	81.35	5	2765.90	1	
2	10134	41	94.74	2	3884.34	1	
3	10145	45	83.26	6	3746.70	1	
5	10168	36	96.66	1	3479.76	1	

	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE	MSRP	DEALSIZE
0	1	2	2003	0	95	0
1	2	5	2003	0	95	0
2	3	7	2003	0	95	1

```
3       3          8     2003            0    95          1
5       4         10     2003            0    95          1
```

[ ]:
```
X = sds.iloc[:,:-1].values #rows and then columns in brackets
Y = sds.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[1.0107e+04, 3.0000e+01, 9.5700e+01, …, 2.0030e+03, 0.0000e+00,
        9.5000e+01],
       [1.0121e+04, 3.4000e+01, 8.1350e+01, …, 2.0030e+03, 0.0000e+00,
        9.5000e+01],
       [1.0134e+04, 4.1000e+01, 9.4740e+01, …, 2.0030e+03, 0.0000e+00,
        9.5000e+01],
       …,
       [1.0386e+04, 4.3000e+01, 1.0000e+02, …, 2.0050e+03, 5.0000e+00,
        5.4000e+01],
       [1.0397e+04, 3.4000e+01, 6.2240e+01, …, 2.0050e+03, 5.0000e+00,
        5.4000e+01],
       [1.0414e+04, 4.7000e+01, 6.5520e+01, …, 2.0050e+03, 5.0000e+00,
        5.4000e+01]])
```

[ ]:
```
Y
```

[ ]:
```
array([0, 0, 1, …, 1, 0, 1], dtype=int64)
```

1.7.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
# assignment what is random_state=42
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=42)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
        # assignment which one we should accept from these
```

```
LinearRegression()
R2 score is:  0.7219737263959292
Mean Absolute error is:  0.2567892880309128
Mean Squared Error:  0.08879286038550639
---------------------------------------------------

C:\Users\Anonymous\AppData\Local\Programs\Python\Python310\lib\site-
packages\sklearn\linear_model\_logistic.py:814: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(

LogisticRegression()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
---------------------------------------------------
GaussianNB()
R2 score is:  0.6470308149344264
Mean Absolute error is:  0.11272727272727273
Mean Squared Error:  0.11272727272727273
---------------------------------------------------
RandomForestClassifier()
R2 score is:  1.0
Mean Absolute error is:  0.0
```

```
Mean Squared Error:  0.0
--------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
--------------------------------------------------
SVR()
R2 score is:  0.7270045939862784
Mean Absolute error is:  0.25023766264365555
Mean Squared Error:  0.08718615927133762
--------------------------------------------------
KNeighborsRegressor()
R2 score is:  0.9936237824633316
Mean Absolute error is:  0.005090909090909091
Mean Squared Error:  0.0020363636363636365
--------------------------------------------------
```

1.7.3 EDA On Winehq_Bug_Report_Data

[ ]:
```
wbrd.head()
```

[ ]:
```
       bug_id creation_date component_name product_name  \
0   WINE-5930    2006-08-12    directx-d3d         WINE
1  WINE-28497    2011-09-25          ntdll         WINE
2  WINE-17067    2009-01-21       -unknown         WINE
3  WINE-42573    2017-03-01       -unknown         WINE
4  WINE-34327    2013-08-22       -unknown         WINE

                                 short_description  \
0                 Graphic glitches in Alien Shooter
1           Dawn of War: Soulstorm no longer starts
2               Stud_PE crashes on Tools -> Plugins
3   Several Flickering Senran Kagura Shinovi Versus
4  Adobe Acrobat X Pro/Standard installer fails t…

                                  long_description assignee_name  \
0  After initial loading screen ingame menu shoul…     wine-bugs
1  After upgrading my wine to 1.3.29 Soulstorm no…     wine-bugs
2  Stud_PE 4.2.0.1 the EXE viewer crashes when th…     wine-bugs
3  Created attachment 57485\nBacktrace-D3D9-Senra…     wine-bugs
4  err:msi:cabinet_copy_file failed to create LC:…     wine-bugs

   reporter_name resolution_category  resolution_code status_category  \
0         karaluh               fixed               1          closed
1            erik               fixed               1          closed
2        specious               fixed               1          closed
```

```
3    mrdeathjr28              fixed                1          closed
4  neilhellfeldt              fixed                1          closed

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            6  2008-10-24                  0                    19
1            6  2011-10-21                  0                    16
2            6  2010-06-11                  0                     8
3            6  2019-02-15                  0                     6
4            6  2013-12-20                  0                     5

  resolution_date  bug_fix_time severity_category  severity_code
0      2008-10-24           804            normal              2
1      2011-10-12            17            normal              2
2      2010-06-03           498            normal              2
3      2019-02-11           712            normal              2
4      2013-12-07           107            normal              2
```

[ ]:
```
wbrd.tail()
```

[ ]:
```
          bug_id creation_date component_name product_name  \
6069  WINE-19917    2009-09-02       -unknown         WINE
6070  WINE-36757    2014-06-17         scrrun         WINE
6071  WINE-13280    2008-05-17        wininet         WINE
6072  WINE-41596    2016-10-23    directx-d3d         WINE
6073   WINE-4375    2006-01-22       -unknown         WINE

                                  short_description  \
6069  ABBYY Lingvo x3/x5 English Edition: crash when…
6070  Microsoft Visual Studio 2005 Express reports '…
6071                      Klipfolio 4 hangs on startup
6072                      Insane 2 crashes in the menu
6073                                   corrupt wine

                                   long_description assignee_name  \
6069  Created attachment 23384\nGeneral wine console…     wine-bugs
6070  Hello folks\n\nas the summary says.\n\nCreate …     wine-bugs
6071  Created attachment 13129\nKlipfolio 4 logs\n\n…     wine-bugs
6072  Created attachment 55938\nterminal output\n\nT…     wine-bugs
6073  I installed wine a couple of days ago and it w…     wine-bugs

        reporter_name resolution_category  resolution_code status_category  \
6069  ihar.hrachyshka               fixed               1          closed
6070            focht               fixed               1          closed
6071          nodisgod              fixed               1          closed
6072          gyebro69              fixed               1          closed
6073   cgoelectronics               fixed               1          closed
```

```
      status_code update_date  quantity_of_votes  quantity_of_comments  \
6069            6  2013-06-07                  0                    14
6070            6  2014-06-27                  0                     4
6071            6  2009-01-13                  0                     7
6072            6  2016-11-11                  0                     7
6073            6  2016-01-05                  0                     5

     resolution_date  bug_fix_time severity_category  severity_code
6069      2013-05-27          1363            normal              2
6070      2014-06-23             6            normal              2
6071      2008-06-02            16            normal              2
6072      2016-10-31             8            normal              2
6073      2006-01-22             0            normal              2
```

[ ]:
```
wbrd.describe()
```

[ ]:
```
       resolution_code  status_code  quantity_of_votes  quantity_of_comments  \
count           6074.0  6074.000000             6074.0           6074.000000
mean               1.0     5.999012                0.0             11.245473
std                0.0     0.044441                0.0             13.275749
min                1.0     4.000000                0.0              0.000000
25%                1.0     6.000000                0.0              5.000000
50%                1.0     6.000000                0.0              8.000000
75%                1.0     6.000000                0.0             13.000000
max                1.0     6.000000                0.0            439.000000

       bug_fix_time  severity_code
count   6074.000000    6074.000000
mean     490.597300       2.080507
std      665.082712       0.594673
min        0.000000       1.000000
25%       22.000000       2.000000
50%      220.000000       2.000000
75%      708.750000       2.000000
max     5636.000000       6.000000
```

[ ]:
```
wbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6074 entries, 0 to 6073
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                6074 non-null   object
 1   creation_date         6074 non-null   object
 2   component_name        6074 non-null   object
 3   product_name          6074 non-null   object
```

```
 4   short_description     6074 non-null   object
 5   long_description      6063 non-null   object
 6   assignee_name         6074 non-null   object
 7   reporter_name         6074 non-null   object
 8   resolution_category   6074 non-null   object
 9   resolution_code       6074 non-null   int64
 10  status_category       6074 non-null   object
 11  status_code           6074 non-null   int64
 12  update_date           6074 non-null   object
 13  quantity_of_votes     6074 non-null   int64
 14  quantity_of_comments  6074 non-null   int64
 15  resolution_date       6074 non-null   object
 16  bug_fix_time          6074 non-null   int64
 17  severity_category     6074 non-null   object
 18  severity_code         6074 non-null   int64
dtypes: int64(6), object(13)
memory usage: 901.7+ KB
```

[ ]:
```
wbrd['status_category'].unique()
```

[ ]:
```
array(['closed', 'resolved'], dtype=object)
```

[ ]:
```
# As resolution category has only one element so we wont touch it
# status_category we can encode it as it is two categories only
# We are encoding it
wbrd['status_category'] = wbrd['status_category'].replace("resolved", 1)
wbrd['status_category'] = wbrd['status_category'].replace("closed", 0)
```

[ ]:
```
wbrd['severity_category'].unique()
```

[ ]:
```
array(['normal', 'minor', 'trivial', 'critical', 'major', 'blocker'],
      dtype=object)
```

[ ]:
```
# As we have another column severity_category and we can perform encoding on it␣
 ↪as well
wbrd['severity_category'] = wbrd['severity_category'].replace("normal", 0)
wbrd['severity_category'] = wbrd['severity_category'].replace("blocker", 1)
wbrd['severity_category'] = wbrd['severity_category'].replace("trivial", 2)
wbrd['severity_category'] = wbrd['severity_category'].replace("minor", 3)
wbrd['severity_category'] = wbrd['severity_category'].replace("major", 4)
wbrd['severity_category'] = wbrd['severity_category'].replace("critical", 5)
```

[ ]:
```
wbrd.head(15)
```

[ ]:
```
        bug_id creation_date component_name product_name  \
0    WINE-5930    2006-08-12    directx-d3d         WINE
1   WINE-28497    2011-09-25          ntdll         WINE
```

```
2   WINE-17067    2009-01-21       -unknown         WINE
3   WINE-42573    2017-03-01       -unknown         WINE
4   WINE-34327    2013-08-22       -unknown         WINE
5   WINE-12895    2008-04-30        winedbg         WINE
6     WINE-709    2002-05-18       richedit         WINE
7    WINE-5673    2006-07-14       -unknown         WINE
8   WINE-33873    2013-06-24       -unknown         WINE
9    WINE-1887    2003-12-15            msi         WINE
10  WINE-29717    2012-01-27       -unknown         WINE
11  WINE-35110    2013-12-11       vbscript         WINE
12   WINE-9987    2007-10-11          gdi32         WINE
13  WINE-12570    2008-04-13       -unknown         WINE
14  WINE-20299    2009-10-09         user32         WINE

                                  short_description  \
0                    Graphic glitches in Alien Shooter
1              Dawn of War: Soulstorm no longer starts
2                  Stud_PE crashes on Tools -> Plugins
3     Several Flickering Senran Kagura Shinovi Versus
4   Adobe Acrobat X Pro/Standard installer fails t…
5          Winedbg can't return the value of a double
6   riched32.dll needs to be improved a lot (was:R…
7         Monkey Island 4 (MI4) demo does not install
8                       Photoshop CS2 fails to save as
9             microsoft movie maker 2 installer fails
10             IE4 setup wants inetcpl.cpl.DllInstall
11  LabChart Reader 8 installer ends prematurely (…
12  Missing function GDI32.dll.RemoveFontMemResour…
13  AutoCAD 2008 Register Today window contents no…
14  AutoCAD 2008: No images on buttons from Quick …

                                    long_description assignee_name  \
0   After initial loading screen ingame menu shoul…     wine-bugs
1   After upgrading my wine to 1.3.29 Soulstorm no…     wine-bugs
2   Stud_PE 4.2.0.1 the EXE viewer crashes when th…     wine-bugs
3   Created attachment 57485\nBacktrace-D3D9-Senra…     wine-bugs
4   err:msi:cabinet_copy_file failed to create LC:…     wine-bugs
5   struct test\n{\n    int a\n    double b\n    i…     wine-bugs
6   PTE is a text editor (you can get this freewar…          wdev
7   The MI4 demo uses InstallShield. Everything se…     wine-bugs
8   Created attachment 44936\nScreenshot\n\nWhen I…     wine-bugs
9   It's a xp only app. The installer starts asks …          mike
10  After rebooting the Internet Explorer 4 setup …     wine-bugs
11  Created attachment 46834\nstandard output\n\nS…     wine-bugs
12  Created attachment 8507\nRemoveFontMemResource…     wine-bugs
13  Created attachment 12150\nlog\n\nI cannot regi…     wine-bugs
14  Created attachment 24000\nQuick Help Toolbar\n…     wine-bugs
```

```
         reporter_name resolution_category  resolution_code  status_category  \
0              karaluh               fixed                1                0
1                 erik               fixed                1                0
2              specious               fixed                1                0
3           mrdeathjr28               fixed                1                0
4         neilhellfeldt               fixed                1                0
5                 wine               fixed                1                0
6               myasar               fixed                1                0
7         daniel.skorka               fixed                1                0
8            jeff.artik               fixed                1                0
9               ivanleo               fixed                1                0
10   RandomAccountName               fixed                1                0
11              00cpxxx               fixed                1                0
12           frans.kool               fixed                1                0
13         dominikowski               fixed                1                0
14  lukasz.wojnilowicz               fixed                1                0

    status_code update_date  quantity_of_votes  quantity_of_comments  \
0             6  2008-10-24                  0                    19
1             6  2011-10-21                  0                    16
2             6  2010-06-11                  0                     8
3             6  2019-02-15                  0                     6
4             6  2013-12-20                  0                     5
5             6  2008-05-09                  0                     4
6             6  2012-02-23                  0                    28
7             6  2007-04-03                  0                     5
8             6  2014-10-31                  0                    37
9             6  2010-04-04                  0                     8
10            6  2012-06-15                  0                     5
11            6  2014-03-21                  0                     7
12            6  2009-01-12                  0                    17
13            6  2012-09-14                  0                    82
14            6  2010-03-19                  0                     8

   resolution_date  bug_fix_time  severity_category  severity_code
0       2008-10-24           804                  0              2
1       2011-10-12            17                  0              2
2       2010-06-03           498                  0              2
3       2019-02-11           712                  0              2
4       2013-12-07           107                  0              2
5       2008-05-03             3                  0              2
6       2008-11-29          2387                  0              2
7       2007-04-03           263                  0              2
8       2014-10-27           490                  0              2
9       2006-01-03           750                  0              2
10      2012-03-13            46                  3              2
```

11	2014-03-08	87	0	2
12	2008-01-02	83	0	2
13	2012-09-09	1610	0	2
14	2010-03-15	157	2	1

[ ]:
```
wbrd['component_name'].unique()
```

[ ]:
```
array(['directx-d3d', 'ntdll', '-unknown', 'winedbg', 'richedit', 'msi',
       'vbscript', 'gdi32', 'user32', 'quartz', 'kernel32', 'gdiplus',
       'wineoss.drv', 'wineserver', 'shdocvw', 'ole32', 'loader',
       'testcases', 'mshtml', 'advapi32', 'cmd', 'urlmon', 'wininet',
       'msvcrt', 'msvcp', 'rpc', 'winex11.drv', 'msxml3', 'comctl32',
       'www-unknown', 'appdb-unknown', 'wmp&wmvcore', 'apphelp',
       'build-env', 'tools', 'ktmw32', 'wineps.drv', 'rsaenh', 'setupapi',
       'oleaut32', 'winsock', 'bugzilla-unknown', 'usp10', 'programs',
       'xaudio2', 'directx-dsound', 'crypt32', 'wevtapi', 'browseui',
       'registry', 'opengl', 'directx-d3dx9', 'directx-d3dx11', 'atl',
       'jscript', 'shell32', 'ndis.sys', 'msacm32', 'unknown', 'dbghelp',
       'directx-dinput', 'winegstreamer', 'wmi&wbemprox', 'hid',
       'winhttp', 'msvcirt', 'mscms', 'pdh', 'd2d', 'winmm&mci',
       'windowscodecs', 'comdlg32', 'ieframe', 'wine-packages', 'uxtheme',
       'documentation', 'slc', 'scrrun', 'mfplat', 'wscript', 'spooler',
       'dwrite', 'mountmgr.sys', 'xinput', 'msxml4', 'directx-dplay',
       'winealsa.drv', 'propsys', 'iphlpapi', 'ntoskrnl',
       'uiautomationcore', 'msasn1', 'msadp32', 'netcfgx', 'mscoree',
       'imagehlp', 'odbc', 'xapofx', 'loadperf', 'wintrust', 'netapi32',
       'mlang', 'msctf', 'secur32', 'winemac.drv', 'kernelbase',
       'directx-dmusic', 'oledb32', 'shlwapi', 'hnetcfg', 'dwmapi', 'hal',
       'l3codeca.acm', 'fonts', 'wshom.ocx', 'directx-d3dx10',
       'directx-d3dxof', 'bcrypt', 'ninput', 'hhctrl.ocx', 'cabinet',
       'mmdevapi', 'ole16', 'msvfw32', 'winepulse.drv', 'vcomp',
       'openal32', 'winsta', 'winscard', 'dos', 'api-ms-win-*', 'mstask',
       'gameux', 'x3daudio', 'msvidc32', 'sfc', 'oleacc', 'taskschd',
       'iccvid', 'fusion', 'netprofm', 'imm32', 'dxva2', 'wia', 'wlanapi',
       'qmgr', 'glu32', 'advpack', 'tapi32', 'dnsapi', 'version'],
      dtype=object)
```

[ ]:
```
wbrd['product_name'].unique()
```

[ ]:
```
array(['WINE', 'WINEHQ.ORG', 'WINEHQ_APPS_DATABASE', 'WINEHQ_BUGZILLA',
       'WINE-TESTBOT', 'PACKAGING', 'WINE-STAGING'], dtype=object)
```

1.8 As other elements are not having a categorical

1.8.1 Now we will be removing null values

[ ]:
```
wbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description       11
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
# percentage of missing values
wbrd.isnull().sum() / wbrd.shape[0] * 100
```

[ ]:
```
bug_id                  0.0000
creation_date           0.0000
component_name          0.0000
product_name            0.0000
short_description       0.0000
long_description        0.1811
assignee_name           0.0000
reporter_name           0.0000
resolution_category     0.0000
resolution_code         0.0000
status_category         0.0000
status_code             0.0000
update_date             0.0000
quantity_of_votes       0.0000
quantity_of_comments    0.0000
resolution_date         0.0000
bug_fix_time            0.0000
severity_category       0.0000
```

```
severity_code           0.0000
dtype: float64
```

[ ]:
```
wbrd.dropna(inplace=True)
```

[ ]:
```
wbrd.isnull().sum()
```

[ ]:
```
bug_id                  0
creation_date           0
component_name          0
product_name            0
short_description       0
long_description        0
assignee_name           0
reporter_name           0
resolution_category     0
resolution_code         0
status_category         0
status_code             0
update_date             0
quantity_of_votes       0
quantity_of_comments    0
resolution_date         0
bug_fix_time            0
severity_category       0
severity_code           0
dtype: int64
```

[ ]:
```
wbrd.head()
```

[ ]:
```
       bug_id creation_date component_name product_name  \
0   WINE-5930    2006-08-12    directx-d3d         WINE
1  WINE-28497    2011-09-25          ntdll         WINE
2  WINE-17067    2009-01-21       -unknown         WINE
3  WINE-42573    2017-03-01       -unknown         WINE
4  WINE-34327    2013-08-22       -unknown         WINE

                                 short_description  \
0                 Graphic glitches in Alien Shooter
1           Dawn of War: Soulstorm no longer starts
2               Stud_PE crashes on Tools -> Plugins
3   Several Flickering Senran Kagura Shinovi Versus
4  Adobe Acrobat X Pro/Standard installer fails t…

                                  long_description assignee_name  \
0  After initial loading screen ingame menu shoul…     wine-bugs
1  After upgrading my wine to 1.3.29 Soulstorm no…     wine-bugs
```

```
2  Stud_PE 4.2.0.1 the EXE viewer crashes when th…     wine-bugs
3  Created attachment 57485\nBacktrace-D3D9-Senra…     wine-bugs
4  err:msi:cabinet_copy_file failed to create LC:…     wine-bugs

   reporter_name resolution_category  resolution_code  status_category  \
0         karaluh               fixed               1                0
1            erik               fixed               1                0
2        specious               fixed               1                0
3     mrdeathjr28               fixed               1                0
4   neilhellfeldt               fixed               1                0

   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            6  2008-10-24                  0                    19
1            6  2011-10-21                  0                    16
2            6  2010-06-11                  0                     8
3            6  2019-02-15                  0                     6
4            6  2013-12-20                  0                     5

  resolution_date  bug_fix_time  severity_category  severity_code
0      2008-10-24           804                  0              2
1      2011-10-12            17                  0              2
2      2010-06-03           498                  0              2
3      2019-02-11           712                  0              2
4      2013-12-07           107                  0              2
```

[ ]:
```
from datetime import time,date,datetime

wbrd['creation_date'] = pd.to_datetime(wbrd['creation_date'])
wbrd['update_date'] = pd.to_datetime(wbrd['update_date'])
wbrd['resolution_date'] = pd.to_datetime(wbrd['resolution_date'])
```

[ ]:
```
wbrd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6063 entries, 0 to 6073
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   bug_id                6063 non-null   object
 1   creation_date         6063 non-null   datetime64[ns]
 2   component_name        6063 non-null   object
 3   product_name          6063 non-null   object
 4   short_description     6063 non-null   object
 5   long_description      6063 non-null   object
 6   assignee_name         6063 non-null   object
 7   reporter_name         6063 non-null   object
 8   resolution_category   6063 non-null   object
```

```
 9   resolution_code       6063 non-null   int64
 10  status_category       6063 non-null   int64
 11  status_code           6063 non-null   int64
 12  update_date           6063 non-null   datetime64[ns]
 13  quantity_of_votes     6063 non-null   int64
 14  quantity_of_comments  6063 non-null   int64
 15  resolution_date       6063 non-null   datetime64[ns]
 16  bug_fix_time          6063 non-null   int64
 17  severity_category     6063 non-null   int64
 18  severity_code         6063 non-null   int64
dtypes: datetime64[ns](3), int64(8), object(8)
memory usage: 947.3+ KB
```

[ ]:
```
wbrd.insert(0, "time_to_dep[s]", ((
    wbrd['resolution_date']-wbrd['creation_date']).astype('timedelta64[s]')),␣
 ↪True)
```

[ ]:
```
wbrd.head()
```

[ ]:
```
   time_to_dep[s]      bug_id creation_date component_name product_name  \
0      69465600.0   WINE-5930    2006-08-12    directx-d3d         WINE
1       1468800.0  WINE-28497    2011-09-25          ntdll         WINE
2      43027200.0  WINE-17067    2009-01-21       -unknown         WINE
3      61516800.0  WINE-42573    2017-03-01       -unknown         WINE
4       9244800.0  WINE-34327    2013-08-22       -unknown         WINE

                                short_description  \
0                 Graphic glitches in Alien Shooter
1           Dawn of War: Soulstorm no longer starts
2               Stud_PE crashes on Tools -> Plugins
3   Several Flickering Senran Kagura Shinovi Versus
4  Adobe Acrobat X Pro/Standard installer fails t…

                                 long_description assignee_name  \
0  After initial loading screen ingame menu shoul…     wine-bugs
1  After upgrading my wine to 1.3.29 Soulstorm no…     wine-bugs
2  Stud_PE 4.2.0.1 the EXE viewer crashes when th…     wine-bugs
3  Created attachment 57485\nBacktrace-D3D9-Senra…     wine-bugs
4  err:msi:cabinet_copy_file failed to create LC:…     wine-bugs

   reporter_name resolution_category  resolution_code  status_category  \
0         karaluh               fixed               1                0
1            erik               fixed               1                0
2        specious               fixed               1                0
3     mrdeathjr28               fixed               1                0
4  neilhellfeldt               fixed               1                0
```

```
   status_code update_date  quantity_of_votes  quantity_of_comments  \
0            6  2008-10-24                  0                    19
1            6  2011-10-21                  0                    16
2            6  2010-06-11                  0                     8
3            6  2019-02-15                  0                     6
4            6  2013-12-20                  0                     5

  resolution_date  bug_fix_time  severity_category  severity_code
0      2008-10-24           804                  0              2
1      2011-10-12            17                  0              2
2      2010-06-03           498                  0              2
3      2019-02-11           712                  0              2
4      2013-12-07           107                  0              2
```

[ ]:
```
# dropping extra columns
wbrd.
 ↪drop(['bug_id','resolution_code','quantity_of_votes','creation_date','component_name','prod
```

[ ]:
```
# now we are done with dropping values as well
wbrd['bug_fix_time'].agg(['skew', 'kurtosis']).transpose()
```

[ ]:
```
skew        2.140220
kurtosis    5.540213
Name: bug_fix_time, dtype: float64
```

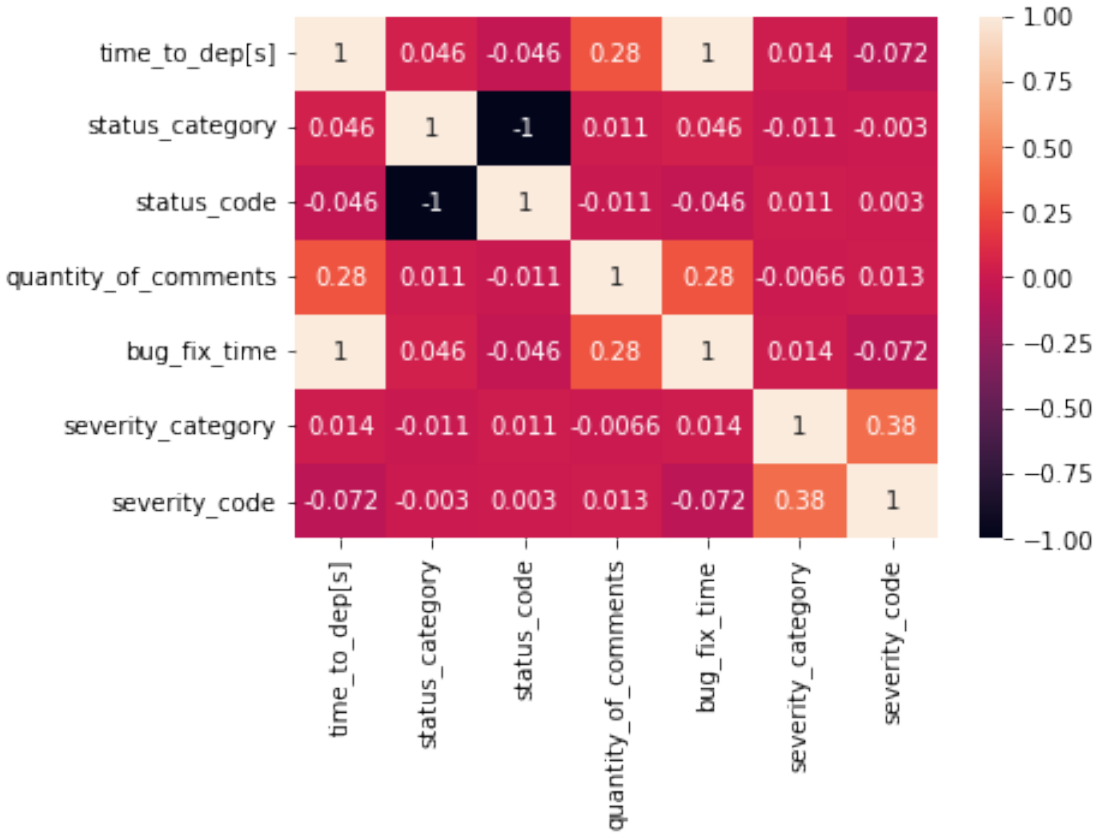
[ ]:
```
corr = wbrd.corr(method="pearson")  # you can use spearman if you want
corr
```

[ ]:
```
                      time_to_dep[s]  status_category  status_code  \
time_to_dep[s]              1.000000         0.046059    -0.046059
status_category             0.046059         1.000000    -1.000000
status_code                -0.046059        -1.000000     1.000000
quantity_of_comments        0.282762         0.011286    -0.011286
bug_fix_time                1.000000         0.046059    -0.046059
severity_category           0.014405        -0.010645     0.010645
severity_code              -0.072357        -0.002983     0.002983

                      quantity_of_comments  bug_fix_time  severity_category  \
time_to_dep[s]                    0.282762      1.000000           0.014405
status_category                   0.011286      0.046059          -0.010645
status_code                      -0.011286     -0.046059           0.010645
quantity_of_comments              1.000000      0.282762          -0.006614
bug_fix_time                      0.282762      1.000000           0.014405
severity_category                -0.006614      0.014405           1.000000
severity_code                     0.012688     -0.072357           0.383389

                      severity_code
```

```
time_to_dep[s]          -0.072357
status_category         -0.002983
status_code              0.002983
quantity_of_comments     0.012688
bug_fix_time            -0.072357
severity_category        0.383389
severity_code            1.000000
```

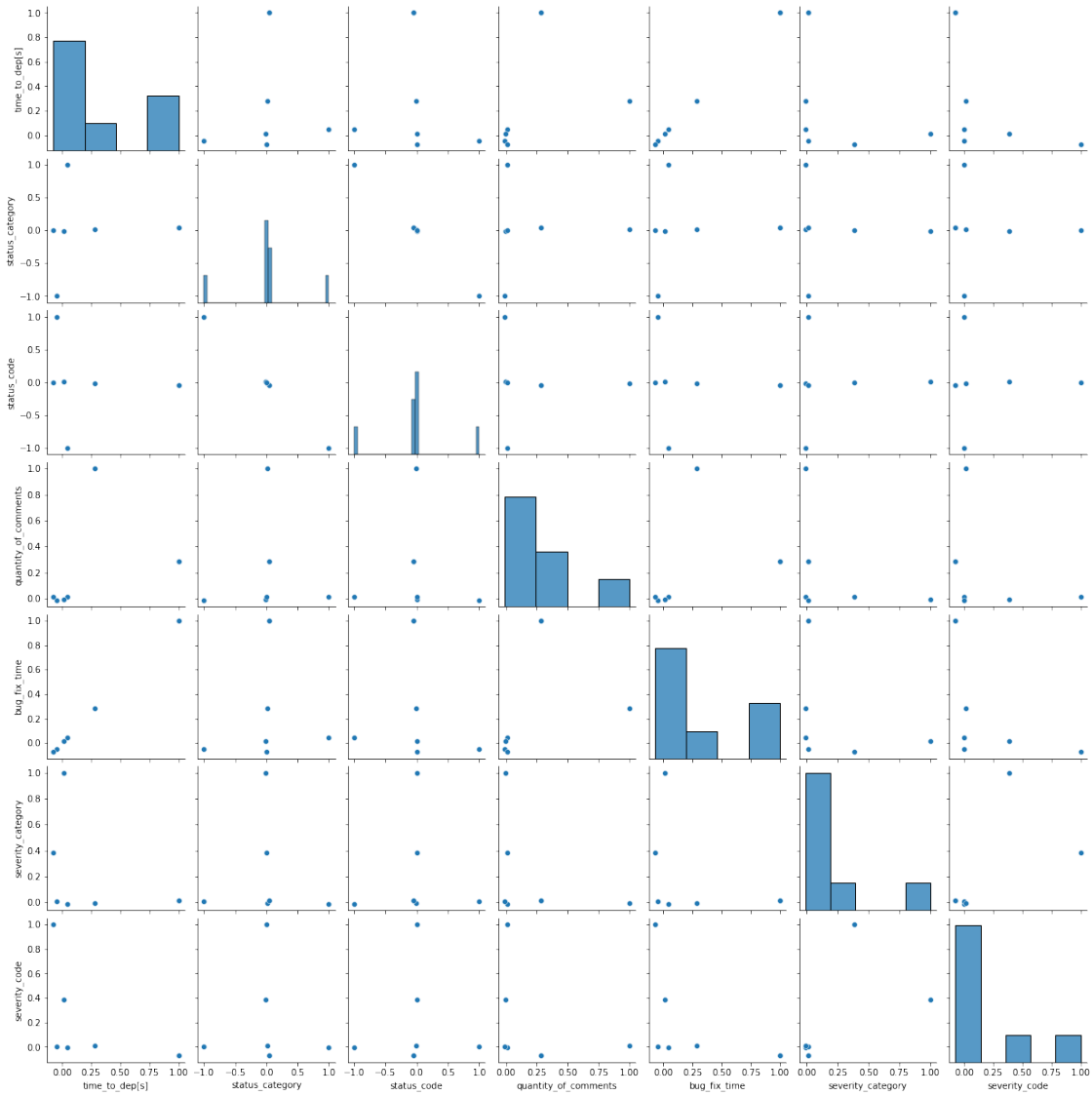
[ ]:
```
sns.heatmap(corr, annot=True)
```

[ ]:
```
<AxesSubplot:>
```



[ ]:
```
# we can also draw a pairplot to see the correlation
sns.pairplot(corr)
```

[ ]:
```
<seaborn.axisgrid.PairGrid at 0x266c0444cd0>
```



[ ]:
```
wbrd.head()
```

[ ]:

	time_to_dep[s]	status_category	status_code	quantity_of_comments	\
0	69465600.0	0	6	19	
1	1468800.0	0	6	16	
2	43027200.0	0	6	8	
3	61516800.0	0	6	6	
4	9244800.0	0	6	5	

	bug_fix_time	severity_category	severity_code
0	804	0	2
1	17	0	2
2	498	0	2

```
3          712                0             2
4          107                0             2
```

[ ]:
```
X = wbrd.iloc[:,:-1].values #rows and then columns in brackets
Y = wbrd.iloc[:,-1].values
```

[ ]:
```
X
```

[ ]:
```
array([[6.94656e+07, 0.00000e+00, 6.00000e+00, 1.90000e+01, 8.04000e+02,
        0.00000e+00],
       [1.46880e+06, 0.00000e+00, 6.00000e+00, 1.60000e+01, 1.70000e+01,
        0.00000e+00],
       [4.30272e+07, 0.00000e+00, 6.00000e+00, 8.00000e+00, 4.98000e+02,
        0.00000e+00],
       …,
       [1.38240e+06, 0.00000e+00, 6.00000e+00, 7.00000e+00, 1.60000e+01,
        0.00000e+00],
       [6.91200e+05, 0.00000e+00, 6.00000e+00, 7.00000e+00, 8.00000e+00,
        0.00000e+00],
       [0.00000e+00, 0.00000e+00, 6.00000e+00, 5.00000e+00, 0.00000e+00,
        0.00000e+00]])
```

[ ]:
```
Y
```

[ ]:
```
array([2, 2, 2, …, 2, 2, 2], dtype=int64)
```

1.8.2 Training data

[ ]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

[ ]:
```
lr = LinearRegression()
lrr = LogisticRegression()
nb = GaussianNB()
rf = RandomForestClassifier()
dt = DecisionTreeRegressor()
svr = SVR()
krn = KNeighborsRegressor()
```

[ ]:
```
# model loop
# assignment what is random_state=42
X_train,X_test,y_train,y_test =  train_test_split(X,Y,test_size=0.
 ↪2,random_state=42)
for i in [lr,lrr,nb,rf,dt,svr,krn]: # read all models
    i.fit(X_train,y_train) # fitting our models
    pred= i.predict(X_test) # predict
    test_score = r2_score(y_test,pred) # test_score
    train_score = r2_score(y_train,i.predict(X_train)) # train score
    if abs(train_score-test_score <= 0.1):
        print(i)
        print('R2 score is: ', r2_score(y_test,pred))
        print('Mean Absolute error is: ', mean_absolute_error(y_test, pred))
        print('Mean Squared Error: ', mean_squared_error(y_test,pred))
        print("---------------------------------------------------")
        # assignment which one we should accept from these
```

```
LinearRegression()
R2 score is:  0.16588550909858812
Mean Absolute error is:  0.21938897121242598
Mean Squared Error:  0.2781320795497475
---------------------------------------------------
LogisticRegression()
R2 score is:  -0.016144812095666428
Mean Absolute error is:  0.1508656224237428
Mean Squared Error:  0.3388293487221764
---------------------------------------------------
GaussianNB()
R2 score is:  -0.02108955570683757
Mean Absolute error is:  0.15251442704039572
Mean Squared Error:  0.34047815333882936
---------------------------------------------------
RandomForestClassifier()
R2 score is:  0.9381907048603609
Mean Absolute error is:  0.004122011541632316
Mean Squared Error:  0.020610057708161583
---------------------------------------------------
DecisionTreeRegressor()
R2 score is:  1.0
Mean Absolute error is:  0.0
Mean Squared Error:  0.0
---------------------------------------------------
SVR()
R2 score is:  0.002962120857455286
Mean Absolute error is:  0.2294918248155023
Mean Squared Error:  0.33245821975362644
---------------------------------------------------
```

[ ]:

[ ]:
```
#df['gender'] = df['gender'].replace("Male", 1)
#df['gender'] = df['gender'].replace("Female", 0)
```