In [1]:

```
#line1
sc.install_pypi_package("pandas==1.0.3")
sc.install_pypi_package("numpy")
sc.install_pypi_package("matplotlib")
sc.install_pypi_package("matplotlib==3.2.1")
sc.install_pypi_package("seaborn==0.10.0")
```

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1638463447423_0001	pyspark	idle	Link	Link	✓

```
SparkSession available as 'spark'.

Collecting pandas==1.0.3
  Downloading https://files.pythonhosted.org/packages/4a/6a/94b219b8ea0f2d580169e85ed1edc0163743f55aaeca8a44c2e8fc1e344e/
pandas-1.0.3-cp37-cp37m-manylinux1_x86_64.whl (10.0MB)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/site-packages (from pandas==1.0.3)
Requirement already satisfied: numpy>=1.13.3 in /usr/local/lib64/python3.7/site-packages (from pandas==1.0.3)
Collecting python-dateutil>=2.6.1 (from pandas==1.0.3)
  Downloading https://files.pythonhosted.org/packages/36/7a/87837f39d0296e723bb9b62bbb257d0355c7f6128853c78955f57342a56d/
python_dateutil-2.8.2-py2.py3-none-any.whl (247kB)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil>=2.6.1->pandas==
1.0.3)
Installing collected packages: python-dateutil, pandas
Successfully installed pandas-1.0.3 python-dateutil-2.8.2

Requirement already satisfied: numpy in /usr/local/lib64/python3.7/site-packages


Collecting matplotlib
  Downloading https://files.pythonhosted.org/packages/6b/48/710ebe39563c5f0cb464c6f1c0eb7bfeb6171a151d65ae8254e12306a210/
matplotlib-3.5.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.whl (11.2MB)
Collecting pyparsing>=2.2.1 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/a0/34/895006117f6fce0b4de045c87e154ee4a20c68ec0a4c9a36d900888fb6bc/
pyparsing-3.0.6-py3-none-any.whl (97kB)
Collecting packaging>=20.0 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/05/8e/8de486cbd03baba4deef4142bd643a3e7bbe954a784dc1bb17142572d127/
packaging-21.3-py3-none-any.whl (40kB)
Requirement already satisfied: python-dateutil>=2.7 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from matplot
lib)
Collecting pillow>=6.2.0 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/7d/2a/2fc11b54e2742db06297f7fa7f420a0e3069fdcf0e4b57dfec33f0b08622/
Pillow-8.4.0.tar.gz (49.4MB)
```

```
Collecting cycler>=0.10 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/5c/f9/695d6bedebd747e5eb0fe8fad57b72fdf25411273a39791cde838d5a8f51/
cycler-0.11.0-py3-none-any.whl
Collecting fonttools>=4.22.0 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/f8/a6/6cebbbee2ef1b68022c164192aa232af5f859f3d96126da6ca9f084bce63/
fonttools-4.28.2-py3-none-any.whl (880kB)
Collecting setuptools-scm>=4 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/bc/bf/353180314d0e27929703faf240c244f25ae765e01f595a010cafb209ab51/
setuptools_scm-6.3.2-py3-none-any.whl
Collecting kiwisolver>=1.0.1 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/09/6b/6e567cb2e86d4e5939a9233f8734e26021b6a9c1bc4b1edccba236a84cc2/
kiwisolver-1.3.2-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.whl (1.1MB)
Collecting numpy>=1.17 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/fb/48/b0708ebd7718a8933f0d3937513ef8ef2f4f04529f1f66ca86d873043921/
numpy-1.21.4.zip (10.6MB)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil>=2.7->matplotlib)
Collecting tomli>=1.0.0 (from setuptools-scm>=4->matplotlib)
  Downloading https://files.pythonhosted.org/packages/6d/6c/9908d4db66488217c665a9a5744319406e41f3c46fa5929a8886f2fe1090/
tomli-1.2.2-py3-none-any.whl
Requirement already satisfied: setuptools in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from setuptools-scm>=4
->matplotlib)
Building wheels for collected packages: pillow, numpy
  Running setup.py bdist_wheel for pillow: started
  Running setup.py bdist_wheel for pillow: finished with status 'error'
  Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-
build-aabswi7e/pillow/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close
();exec(compile(code, __file__, 'exec'))" bdist_wheel -d /tmp/tmptqg09zc6pip-wheel- --python-tag cp37:
  /usr/lib64/python3.7/distutils/dist.py:274: UserWarning: Unknown distribution option: 'long_description_content_type'
    warnings.warn(msg)
  /usr/lib64/python3.7/distutils/dist.py:274: UserWarning: Unknown distribution option: 'project_urls'
    warnings.warn(msg)
  running bdist_wheel
  running build
  running build_py
  creating build
  creating build/lib.linux-x86_64-3.7
  creating build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/BdfFontFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/BlpImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/BmpImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/BufrStubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ContainerIO.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/CurImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/DcxImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/DdsImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/EpsImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ExifTags.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/FitsStubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/FliImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/FontFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/FpxImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/FtexImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/GbrImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/GdImageFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/GifImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/GimpGradientFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/GimpPaletteFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/GribStubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/Hdf5StubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/IcnsImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/IcoImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/Image.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageChops.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageCms.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageColor.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageDraw.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageDraw2.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageEnhance.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageFilter.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageFont.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageGrab.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageMath.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageMode.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageMorph.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageOps.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImagePalette.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImagePath.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageQt.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageSequence.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageShow.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageStat.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageTk.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageTransform.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImageWin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/ImtImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/IptcImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/Jpeg2KImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/JpegImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/JpegPresets.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/McIdasImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/MicImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/MpegImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/MpoImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/MspImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PSDraw.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PaletteFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PalmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PcdImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PcfFontFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PcxImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PdfImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PdfParser.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PixarImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PngImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PpmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PsdImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/PyAccess.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/SgiImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/SpiderImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/SunImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/TarIO.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/TgaImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/TiffImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/TiffTags.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/WalImageFile.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/WebPImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/WmfImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/XVThumbImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/XbmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/XpmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/__init__.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/__main__.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/_binary.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/_tkinter_finder.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/_util.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/_version.py -> build/lib.linux-x86_64-3.7/PIL
  copying src/PIL/features.py -> build/lib.linux-x86_64-3.7/PIL
  running egg_info
  writing src/Pillow.egg-info/PKG-INFO
  writing dependency_links to src/Pillow.egg-info/dependency_links.txt
  writing top-level names to src/Pillow.egg-info/top_level.txt
  warning: manifest_maker: standard file '-c' not found

  reading manifest file 'src/Pillow.egg-info/SOURCES.txt'
  reading manifest template 'MANIFEST.in'
  warning: no files found matching '*.c'
```

```
warning: no files found matching '*.h'
warning: no files found matching '*.sh'
warning: no previously-included files found matching '.appveyor.yml'
warning: no previously-included files found matching '.clang-format'
warning: no previously-included files found matching '.coveragerc'
warning: no previously-included files found matching '.editorconfig'
warning: no previously-included files found matching '.readthedocs.yml'
warning: no previously-included files found matching 'codecov.yml'
warning: no previously-included files matching '.git*' found anywhere in distribution
warning: no previously-included files matching '*.pyc' found anywhere in distribution
warning: no previously-included files matching '*.so' found anywhere in distribution
no previously-included directories found matching '.ci'
writing manifest file 'src/Pillow.egg-info/SOURCES.txt'
running build_ext


The headers or library files could not be found for jpeg,
a required dependency when compiling Pillow from source.

Please see the install instructions at:
   https://pillow.readthedocs.io/en/latest/installation.html

Traceback (most recent call last):
  File "/mnt/tmp/pip-build-aabswi7e/pillow/setup.py", line 1024, in <module>
    zip_safe=not (debug_build() or PLATFORM_MINGW),
  File "/usr/lib64/python3.7/distutils/core.py", line 148, in setup
    dist.run_commands()
  File "/usr/lib64/python3.7/distutils/dist.py", line 966, in run_commands
    self.run_command(cmd)
  File "/usr/lib64/python3.7/distutils/dist.py", line 985, in run_command
    cmd_obj.run()
  File "/tmp/1638463975745-0/lib/python3.7/site-packages/wheel/bdist_wheel.py", line 179, in run
    self.run_command('build')
  File "/usr/lib64/python3.7/distutils/cmd.py", line 313, in run_command
    self.distribution.run_command(command)
  File "/usr/lib64/python3.7/distutils/dist.py", line 985, in run_command
    cmd_obj.run()
  File "/usr/lib64/python3.7/distutils/command/build.py", line 135, in run
    self.run_command(cmd_name)
  File "/usr/lib64/python3.7/distutils/cmd.py", line 313, in run_command
    self.distribution.run_command(command)
  File "/usr/lib64/python3.7/distutils/dist.py", line 985, in run_command
    cmd_obj.run()
  File "/tmp/1638463975745-0/lib/python3.7/site-packages/setuptools/command/build_ext.py", line 75, in run
    _build_ext.run(self)
  File "/usr/lib64/python3.7/distutils/command/build_ext.py", line 340, in run
```

```
      self.build_extensions()
    File "/mnt/tmp/pip-build-aabswi7e/pillow/setup.py", line 790, in build_extensions
      raise RequiredDependencyException(f)
  __main__.RequiredDependencyException: jpeg

  During handling of the above exception, another exception occurred:

  Traceback (most recent call last):
    File "<string>", line 1, in <module>
    File "/mnt/tmp/pip-build-aabswi7e/pillow/setup.py", line 1037, in <module>
      raise RequiredDependencyException(msg)
  __main__.RequiredDependencyException:

  The headers or library files could not be found for jpeg,
  a required dependency when compiling Pillow from source.

  Please see the install instructions at:
     https://pillow.readthedocs.io/en/latest/installation.html



  ----------------------------------------
  Running setup.py clean for pillow
  Running setup.py bdist_wheel for numpy: started
  Running setup.py bdist_wheel for numpy: finished with status 'error'
  Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-
build-aabswi7e/numpy/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close();
exec(compile(code, __file__, 'exec'))" bdist_wheel -d /tmp/tmp1nh_k3otpip-wheel- --python-tag cp37:
  Running from numpy source directory.
  Cythonizing sources
  Processing numpy/random/_bounded_integers.pxd.in
  Processing numpy/random/_mt19937.pyx
  Traceback (most recent call last):
    File "/mnt/tmp/pip-build-aabswi7e/numpy/tools/cythonize.py", line 59, in process_pyx
      import Cython
  ModuleNotFoundError: No module named 'Cython'

  The above exception was the direct cause of the following exception:

  Traceback (most recent call last):
    File "/mnt/tmp/pip-build-aabswi7e/numpy/tools/cythonize.py", line 240, in <module>
      main()
    File "/mnt/tmp/pip-build-aabswi7e/numpy/tools/cythonize.py", line 236, in main
      find_process_files(root_dir)
    File "/mnt/tmp/pip-build-aabswi7e/numpy/tools/cythonize.py", line 227, in find_process_files
      process(root_dir, fromfile, tofile, function, hash_db)
    File "/mnt/tmp/pip-build-aabswi7e/numpy/tools/cythonize.py", line 193, in process
      processor_function(fromfile, tofile)
    File "/mnt/tmp/pip-build-aabswi7e/numpy/tools/cythonize.py", line 66, in process_pyx
      raise OSError(msg) from e
  OSError: Cython needs to be installed in Python as a module
  Traceback (most recent call last):
    File "<string>", line 1, in <module>
    File "/mnt/tmp/pip-build-aabswi7e/numpy/setup.py", line 448, in <module>
      setup_package()
    File "/mnt/tmp/pip-build-aabswi7e/numpy/setup.py", line 430, in setup_package
      generate_cython()
    File "/mnt/tmp/pip-build-aabswi7e/numpy/setup.py", line 236, in generate_cython
      raise RuntimeError("Running cythonize failed!")
  RuntimeError: Running cythonize failed!

  ----------------------------------------
  Running setup.py clean for numpy
  Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-
build-aabswi7e/numpy/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close();
exec(compile(code, __file__, 'exec'))" clean --all:
  Running from numpy source directory.

  `setup.py clean` is not supported, use one of the following instead:

    - `git clean -xdf` (cleans all files)
    - `git clean -Xdf` (cleans all versioned files, doesn't touch
                        files that aren't checked into the git repo)

  Add `--force` to your command to use it anyway if you must (unsupported).


  ----------------------------------------
Failed to build pillow numpy
Installing collected packages: pyparsing, packaging, pillow, cycler, fonttools, tomli, setuptools-scm, kiwisolver, numpy,
matplotlib
  Running setup.py install for pillow: started
    Running setup.py install for pillow: finished with status 'error'
    Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pi
p-build-aabswi7e/pillow/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close
();exec(compile(code, __file__, 'exec'))" install --record /tmp/pip-u_azgbmk-record/install-record.txt --single-version-e
xternally-managed --compile --install-headers /tmp/1638463975745-0/include/site/python3.7/pillow:
    /usr/lib64/python3.7/distutils/dist.py:274: UserWarning: Unknown distribution option: 'long_description_content_type'
      warnings.warn(msg)
    /usr/lib64/python3.7/distutils/dist.py:274: UserWarning: Unknown distribution option: 'project_urls'
      warnings.warn(msg)
    running install
    running build
    running build_py
    creating build
    creating build/lib.linux-x86_64-3.7
    creating build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/BdfFontFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/BlpImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/BmpImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/BufrStubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ContainerIO.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/CurImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/DcxImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/DdsImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/EpsImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ExifTags.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/FitsStubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/FliImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/FontFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/FpxImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/FtexImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/GbrImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/GdImageFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/GifImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/GimpGradientFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/GimpPaletteFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/GribStubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/Hdf5StubImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/IcnsImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/IcoImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/Image.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageChops.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageCms.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageColor.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageDraw.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageDraw2.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageEnhance.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageFilter.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageFont.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageGrab.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageMath.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageMode.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageMorph.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageOps.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImagePalette.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImagePath.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageQt.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageSequence.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageShow.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageStat.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageTk.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageTransform.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImageWin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/ImtImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/IptcImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/Jpeg2KImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/JpegImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/JpegPresets.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/McIdasImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/MicImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/MpegImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/MpoImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/MspImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PSDraw.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PaletteFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PalmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PcdImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PcfFontFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PcxImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PdfImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PdfParser.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PixarImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PngImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PpmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PsdImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/PyAccess.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/SgiImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/SpiderImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/SunImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/TarIO.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/TgaImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/TiffImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/TiffTags.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/WalImageFile.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/WebPImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/WmfImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/XVThumbImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/XbmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/XpmImagePlugin.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/__init__.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/__main__.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/_binary.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/_tkinter_finder.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/_util.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/_version.py -> build/lib.linux-x86_64-3.7/PIL
    copying src/PIL/features.py -> build/lib.linux-x86_64-3.7/PIL
    running egg_info
    writing src/Pillow.egg-info/PKG-INFO
    writing dependency_links to src/Pillow.egg-info/dependency_links.txt
    writing top-level names to src/Pillow.egg-info/top_level.txt
    warning: manifest_maker: standard file '-c' not found

    reading manifest file 'src/Pillow.egg-info/SOURCES.txt'
    reading manifest template 'MANIFEST.in'
    warning: no files found matching '*.c'
    warning: no files found matching '*.h'
    warning: no files found matching '*.sh'
    warning: no previously-included files found matching '.appveyor.yml'
    warning: no previously-included files found matching '.clang-format'
    warning: no previously-included files found matching '.coveragerc'
    warning: no previously-included files found matching '.editorconfig'
    warning: no previously-included files found matching '.readthedocs.yml'
    warning: no previously-included files found matching 'codecov.yml'
    warning: no previously-included files matching '.git*' found anywhere in distribution
    warning: no previously-included files matching '*.pyc' found anywhere in distribution
    warning: no previously-included files matching '*.so' found anywhere in distribution
    no previously-included directories found matching '.ci'
    writing manifest file 'src/Pillow.egg-info/SOURCES.txt'
    running build_ext


    The headers or library files could not be found for jpeg,
    a required dependency when compiling Pillow from source.

    Please see the install instructions at:
       https://pillow.readthedocs.io/en/latest/installation.html

    Traceback (most recent call last):
      File "/mnt/tmp/pip-build-aabswi7e/pillow/setup.py", line 1024, in <module>
        zip_safe=not (debug_build() or PLATFORM_MINGW),
      File "/usr/lib64/python3.7/distutils/core.py", line 148, in setup
        dist.run_commands()
      File "/usr/lib64/python3.7/distutils/dist.py", line 966, in run_commands
        self.run_command(cmd)
      File "/usr/lib64/python3.7/distutils/dist.py", line 985, in run_command
        cmd_obj.run()
      File "/tmp/1638463975745-0/lib/python3.7/site-packages/setuptools/command/install.py", line 61, in run
```

```
        return orig.install.run(self)
      File "/usr/lib64/python3.7/distutils/command/install.py", line 556, in run
        self.run_command('build')
      File "/usr/lib64/python3.7/distutils/cmd.py", line 313, in run_command
        self.distribution.run_command(command)
      File "/usr/lib64/python3.7/distutils/dist.py", line 985, in run_command
        cmd_obj.run()
      File "/usr/lib64/python3.7/distutils/command/build.py", line 135, in run
        self.run_command(cmd_name)
      File "/usr/lib64/python3.7/distutils/cmd.py", line 313, in run_command
        self.distribution.run_command(command)
      File "/usr/lib64/python3.7/distutils/dist.py", line 985, in run_command
        cmd_obj.run()
      File "/tmp/1638463975745-0/lib/python3.7/site-packages/setuptools/command/build_ext.py", line 75, in run
        _build_ext.run(self)
      File "/usr/lib64/python3.7/distutils/command/build_ext.py", line 340, in run
        self.build_extensions()
      File "/mnt/tmp/pip-build-aabswi7e/pillow/setup.py", line 790, in build_extensions
        raise RequiredDependencyException(f)
    __main__.RequiredDependencyException: jpeg

    During handling of the above exception, another exception occurred:

    Traceback (most recent call last):
      File "<string>", line 1, in <module>
      File "/mnt/tmp/pip-build-aabswi7e/pillow/setup.py", line 1037, in <module>
        raise RequiredDependencyException(msg)
    __main__.RequiredDependencyException:

    The headers or library files could not be found for jpeg,
    a required dependency when compiling Pillow from source.

    Please see the install instructions at:
       https://pillow.readthedocs.io/en/latest/installation.html



    ----------------------------------------

Collecting matplotlib==3.2.1
  Downloading https://files.pythonhosted.org/packages/b2/c2/71fcf957710f3ba1f09088b35776a799ba7dd95f7c2b195ec800933b276b/
matplotlib-3.2.1-cp37-cp37m-manylinux1_x86_64.whl (12.4MB)
Requirement already satisfied: python-dateutil>=2.1 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from matplot
lib==3.2.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /mnt/tmp/1638463975745-0/lib/python3.7/site-pa
ckages (from matplotlib==3.2.1)
```

```
Collecting cycler>=0.10 (from matplotlib==3.2.1)
  Using cached https://files.pythonhosted.org/packages/5c/f9/695d6bedebd747e5eb0fe8fad57b72fdf25411273a39791cde838d5a8f5
1/cycler-0.11.0-py3-none-any.whl
Requirement already satisfied: numpy>=1.11 in /usr/local/lib64/python3.7/site-packages (from matplotlib==3.2.1)
Collecting kiwisolver>=1.0.1 (from matplotlib==3.2.1)
  Using cached https://files.pythonhosted.org/packages/09/6b/6e567cb2e86d4e5939a9233f8734e26021b6a9c1bc4b1edccba236a84cc
2/kiwisolver-1.3.2-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.whl
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil>=2.1->matplotlib=
=3.2.1)
Installing collected packages: cycler, kiwisolver, matplotlib
Successfully installed cycler-0.11.0 kiwisolver-1.3.2 matplotlib-3.2.1

Collecting seaborn==0.10.0
  Downloading https://files.pythonhosted.org/packages/70/bd/5e6bf595fe6ee0f257ae49336dd180768c1ed3d7c7155b2fdf894c1c808a/
seaborn-0.10.0-py3-none-any.whl (215kB)
Requirement already satisfied: pandas>=0.22.0 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from seaborn==0.1
0.0)
Requirement already satisfied: numpy>=1.13.3 in /usr/local/lib64/python3.7/site-packages (from seaborn==0.10.0)
Collecting scipy>=1.0.1 (from seaborn==0.10.0)
  Downloading https://files.pythonhosted.org/packages/61/67/1a654b96309c991762ee9bc39c363fc618076b155fe52d295211cf2536c7/
scipy-1.7.3.tar.gz (36.1MB)
Requirement already satisfied: matplotlib>=2.1.2 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from seaborn==
0.10.0)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/site-packages (from pandas>=0.22.0->seaborn==0.1
0.0)
Requirement already satisfied: python-dateutil>=2.6.1 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from panda
s>=0.22.0->seaborn==0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /mnt/tmp/1638463975745-0/lib/python3.7/site-pa
ckages (from matplotlib>=2.1.2->seaborn==0.10.0)
Requirement already satisfied: cycler>=0.10 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from matplotlib>=2.
1.2->seaborn==0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /mnt/tmp/1638463975745-0/lib/python3.7/site-packages (from matplotlib
>=2.1.2->seaborn==0.10.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil>=2.6.1->pandas>=
0.22.0->seaborn==0.10.0)
Building wheels for collected packages: scipy
  Running setup.py bdist_wheel for scipy: started
  Running setup.py bdist_wheel for scipy: finished with status 'error'
  Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-
build-o9tjrt7z/scipy/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close();
exec(compile(code, __file__, 'exec'))" bdist_wheel -d /tmp/tmpv43vfvvppip-wheel- --python-tag cp37:
  Error: 'pybind11' must be installed before running the build.


  ----------------------------------------
  Running setup.py clean for scipy
  Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-
build-o9tjrt7z/scipy/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close();
exec(compile(code, __file__, 'exec'))" clean --all:

  `setup.py clean` is not supported, use one of the following instead:

    - `git clean -xdf` (cleans all files)
    - `git clean -Xdf` (cleans all versioned files, doesn't touch
                        files that aren't checked into the git repo)

  Add `--force` to your command to use it anyway if you must (unsupported).


  ----------------------------------------
Failed to build scipy
Installing collected packages: scipy, seaborn
  Running setup.py install for scipy: started
    Running setup.py install for scipy: finished with status 'error'
    Complete output from command /tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pi
p-build-o9tjrt7z/scipy/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close
();exec(compile(code, __file__, 'exec'))" install --record /tmp/pip-7aanzk_r-record/install-record.txt --single-version-e
xternally-managed --compile --install-headers /tmp/1638463975745-0/include/site/python3.7/scipy:

    Note: for reliable uninstall behaviour and dependency installation
    and uninstallation, please use pip instead of using
    `setup.py install`:

      - `pip install .`       (from a git repo or downloaded source
                               release)
      - `pip install scipy`   (last SciPy release on PyPI)


    Error: 'pybind11' must be installed before running the build.


    ----------------------------------------



  Failed building wheel for pillow
  Failed building wheel for numpy
  Failed cleaning build dir for numpy
Command "/tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-build-aabswi7e/pillow/
setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close();exec(compile(code, __
file__, 'exec'))" install --record /tmp/pip-u_azgbmk-record/install-record.txt --single-version-externally-managed --comp
ile --install-headers /tmp/1638463975745-0/include/site/python3.7/pillow" failed with error code 1 in /mnt/tmp/pip-build-
```

```
aabswi7e/pillow/


  Failed building wheel for scipy
  Failed cleaning build dir for scipy
Command "/tmp/1638463975745-0/bin/python -u -c "import setuptools, tokenize;__file__='/mnt/tmp/pip-build-o9tjrt7z/scipy/setup.py';f=getattr(tokenize, 'open', open)(__file__);code=f.read().replace('\r\n', '\n');f.close();exec(compile(code, __file__, 'exec'))" install --record /tmp/pip-7aanzk_r-record/install-record.txt --single-version-externally-managed --compile --install-headers /tmp/1638463975745-0/include/site/python3.7/scipy" failed with error code 1 in /mnt/tmp/pip-build-o9tjrt7z/scipy/
```

In [2]:

```python
#line 2
from pyspark.sql.types import StructType,StructField, StringType, IntegerType, ArrayType
from pyspark.sql.functions import approx_count_distinct
from pyspark.sql.functions import avg
from pyspark.sql.functions import collect_set
from pyspark.sql.functions import countDistinct
from pyspark.sql.functions import count
from pyspark.sql.functions import first, last, max, min
from pyspark.sql.functions import col
from pyspark.sql.functions import array_contains
from pyspark.sql.functions import mean, count, sum, col
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType
from pyspark.sql.functions import explode, split
import pandas as pd
import numpy as np

import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
```

```
unknown magic command 'matplotlib'
UnknownMagic: unknown magic command 'matplotlib'
```

In [3]:

```python
#line 3 和4?
business = spark.read.json('s3://9760project2/project2_yelp/yelp_academic_dataset_business.json')
tip = spark.read.json('s3://9760project2/project2_yelp/yelp_academic_dataset_tip.json')
user = spark.read.json('s3://9760project2/project2_yelp/yelp_academic_dataset_user.json')
checkin= spark.read.json('s3://9760project2/project2_yelp/yelp_academic_dataset_checkin.json')
review = spark.read.json('s3://9760project2/project2_yelp/yelp_academic_dataset_review.json')
```

In [4]:

```
#line5
rows=business.count()
print(rows)
columns = len(business.columns)
print(columns)
```

```
160585
14
```

In [5]:

```
#line7
business_loc =  business.select("business_id", "name","city","state","categories")
business_loc.show(5,truncate= False)
```

```
+----------------------+-----------------------+-----------+-----+-------------------------------------------------------
--------------------------------------------------+
|business_id           |name                   |city       |state|categories
|
+----------------------+-----------------------+-----------+-----+-------------------------------------------------------
--------------------------------------------------+
|6iYb2HFDywm3zjuRg0shjw|Oskar Blues Taproom    |Boulder    |CO   |Gastropubs, Food, Beer Gardens, Restaurants, Bars, Amer
ican (Traditional), Beer Bar, Nightlife, Breweries|
|tCbdrRPZA0oiIYSmHG3J0w|Flying Elephants at PDX|Portland   |OR   |Salad, Soup, Sandwiches, Delis, Restaurants, Cafes, Veg
etarian                                           |
|bvN78flM8NLprQ1a1y5dRg|The Reclaimory         |Portland   |OR   |Antiques, Fashion, Used, Vintage & Consignment, Shoppin
g, Furniture Stores, Home & Garden                |
|oaepsyvc0J17qwi8cfrOWg|Great Clips            |Orange City|FL   |Beauty & Spas, Hair Salons
|
|PE9uqAjdw0E4-8mjGl3wVA|Crossfit Terminus      |Atlanta    |GA   |Gyms, Active Life, Interval Training Gyms, Fitness & In
struction                                         |
+----------------------+-----------------------+-----------+-----+-------------------------------------------------------
--------------------------------------------------+
only showing top 5 rows
```

In [6]:

```
#line 8&9

from pyspark.sql.functions import explode, split
```

```
df=business_loc.select("business_id","categories")
df_exploded = df.withColumn("category",explode(split("categories",","))).drop("categories")
```

In [7]:

```
#line10
df_exploded.show(5,truncate=False)
```

```
+----------------------+-------------+
|business_id           |category     |
+----------------------+-------------+
|6iYb2HFDywm3zjuRg0shjw|Gastropubs   |
|6iYb2HFDywm3zjuRg0shjw| Food        |
|6iYb2HFDywm3zjuRg0shjw| Beer Gardens|
|6iYb2HFDywm3zjuRg0shjw| Restaurants |
|6iYb2HFDywm3zjuRg0shjw| Bars        |
+----------------------+-------------+
only showing top 5 rows
```

In [8]:

```
#line11
df_exploded.select('category').distinct().count()
```

```
2487
```

In [9]:

```
#line12
df_exploded.groupby('category').count().show()
```

```
+-------------------+-----+
|           category|count|
+-------------------+-----+
|     Paddleboarding|   12|
|     Dermatologists|   68|
|              Tires| 1456|
|   Historical Tours|   60|
|              Hakka|    4|
|        Hobby Shops|  135|
|         Bubble Tea|  184|
|            Tanning|  147|
|            Propane|   83|
|           Handyman|   87|
|           Macarons|   50|
|           Japanese| 2039|
| Convenience Stores| 1340|
|        Car Dealers| 1013|
|            Lawyers|  422|
|       IV Hydration|   47|
|            Rolfing|   28|
|            Falafel|   19|
|           Psychics|   62|
|    Tasting Classes|   40|
+-------------------+-----+
only showing top 20 rows
```

In [10]:

```
#line13
bar = df_exploded.groupby('category').count().orderBy("count",ascending=False)
bar.show()
```

```
+--------------------+-----+
|            category|count|
+--------------------+-----+
|         Restaurants|36340|
|                Food|22094|
|            Shopping|20056|
|         Restaurants|14423|
|       Home Services|12001|
|       Beauty & Spas|11633|
|    Health & Medical|11390|
|           Nightlife| 9808|
|      Local Services| 9299|
|                Bars| 8914|
| Event Planning &...| 7617|
|                Food| 7375|
|         Active Life| 7039|
|          Automotive| 6785|
|            Shopping| 6149|
|        Coffee & Tea| 5735|
|          Sandwiches| 5697|
| American (Tradit...| 5235|
|             Fashion| 5231|
|       Beauty & Spas| 4941|
+--------------------+-----+
only showing top 20 rows
```

In [11]:

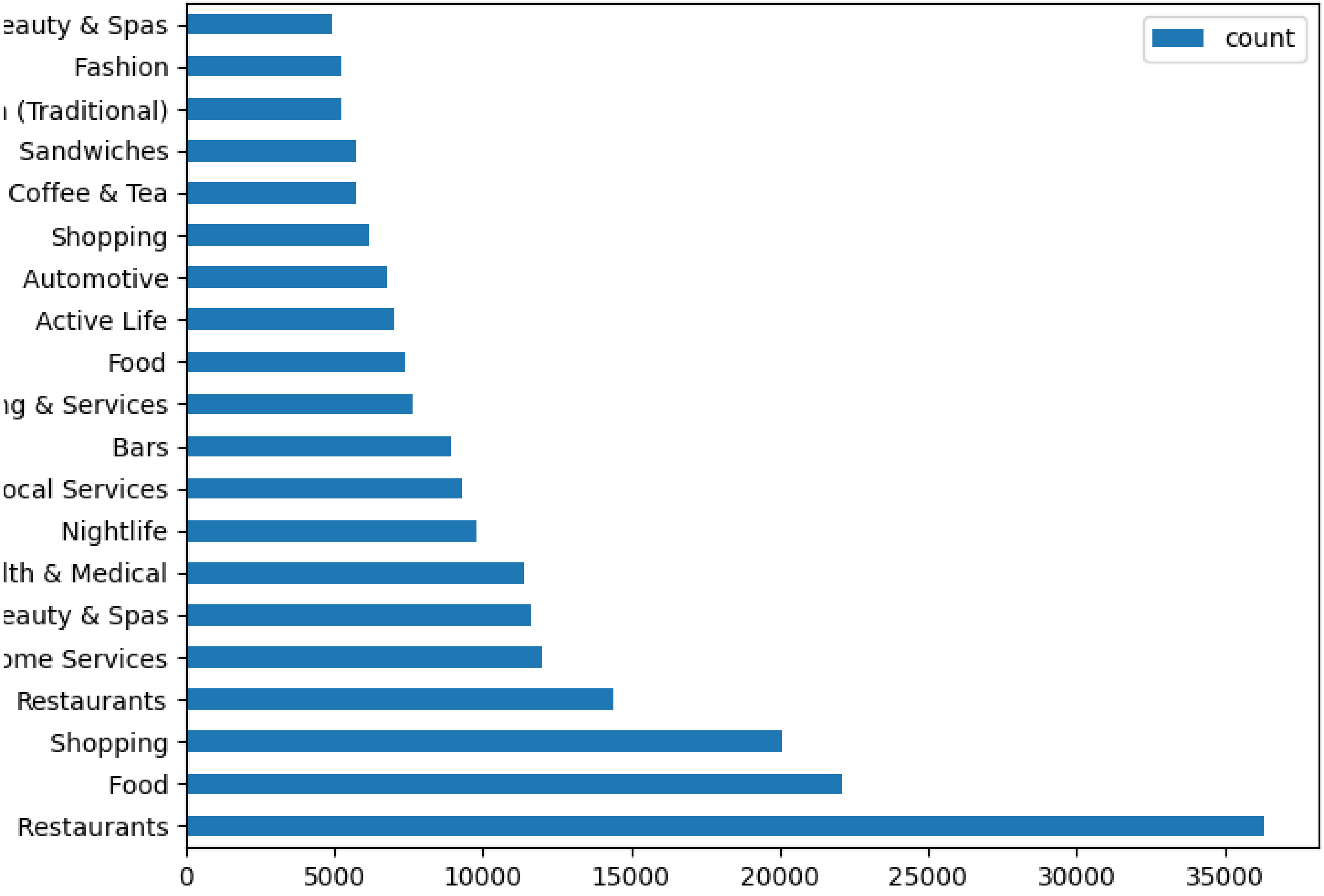
```
#line14
bar_pandas = bar.limit(20).toPandas()
bar_pandas
```

```
                        category  count
0                    Restaurants  36340
1                           Food  22094
2                       Shopping  20056
3                    Restaurants  14423
4                  Home Services  12001
5                  Beauty & Spas  11633
6               Health & Medical  11390
7                      Nightlife   9808
8                 Local Services   9299
9                           Bars   8914
10    Event Planning & Services   7617
11                          Food   7375
12                   Active Life   7039
13                    Automotive   6785
14                      Shopping   6149
15                  Coffee & Tea   5735
16                    Sandwiches   5697
17        American (Traditional)   5235
18                       Fashion   5231
19                 Beauty & Spas   4941
```

In [12]:

```
#line15
plt.figure(figsize=(10,6))
bar_pandas.plot(kind="barh", x="category", figsize=(8,6))

%matplot plt
```



In [13]:

```
#line 16
new_rows=review.count()
print(new_rows)
columns = len(review.columns)
print(columns)
review.printSchema()
```

```
8635403
9
root
 |-- business_id: string (nullable = true)
 |-- cool: long (nullable = true)
 |-- date: string (nullable = true)
 |-- funny: long (nullable = true)
 |-- review_id: string (nullable = true)
 |-- stars: double (nullable = true)
 |-- text: string (nullable = true)
 |-- useful: long (nullable = true)
 |-- user_id: string (nullable = true)
```

In [14]:

```
#line 17 不知道对不对
review =  review.select("business_id","stars")
review.show(5)
```

```
+--------------------+-----+
|         business_id|stars|
+--------------------+-----+
|buF9druCkbuXLX526...|  4.0|
|RA4V8pr014UyUbDvI...|  4.0|
|_sS2LBIGNT5NQb6PD...|  5.0|
|0AzLzHfOJgL7ROwhd...|  2.0|
|8zehGz9jnxPqXtOc7...|  4.0|
+--------------------+-----+
only showing top 5 rows
```

In [15]:

```
#line18
review_dataset= review.groupby("business_id").avg("stars").alias("avg(stars)").sort(col("avg(stars)").desc())
review_dataset.show(5,truncate=False)
```

```
+----------------------+----------+
|business_id           |avg(stars)|
+----------------------+----------+
|4lFIFqycDhV8KQGTFzoEAg|5.0       |
|jZBhw30QecAQNDtpMTSwkA|5.0       |
|_elpnSyXg14LffDRH6HD2w|5.0       |
|j63ReWJhgerb-7em4brmbA|5.0       |
|sxSc6amhnvXbKbaZaqDJLA|5.0       |
+--------------------+----------+
only showing top 5 rows
```

In [16]:

```
#line 19
business_loc= business.select("business_id", "name","city","state","categories","stars")
joined_data = review_dataset.join(business_loc,review_dataset.business_id == business_loc.business_id)#.join(review,revie

joined_data.show(5)
```

```
+--------------------+-----------------+--------------------+--------------------+----------+-----+--------------------+-
----+
|         business_id|       avg(stars)|         business_id|                name|      city|state|          categories|s
tars|
+--------------------+-----------------+--------------------+--------------------+----------+-----+--------------------+-
----+
|R0IJhEI-zSJpYT1YN...|3.606060606060606|R0IJhEI-zSJpYT1YN...|        Nails Studio|    Dedham|   MA|Nail Salons, Waxi...|
3.5|
|wdBrDCbZopowEkIEX...|4.538461538461538|wdBrDCbZopowEkIEX...|Beacon Hill Shoe ...|    Boston|   MA|Local Services, S...|
4.5|
|2boQDeHxopolPtJhV...|4.333333333333333|2boQDeHxopolPtJhV...| Eric Hollander, DDS|    Austin|   TX|Oral Surgeons, En...|
4.5|
|bOnsvrz1VkbrZM1jV...|              3.8|bOnsvrz1VkbrZM1jV...|Fresh Touch Cleaners|Winchester|   MA|Local Services, D...|
4.0|
|XzXcpPCb8Y5huklEN...|4.666666666666667|XzXcpPCb8Y5huklEN...|       Donna & Toots|  Portland|   OR|   Shopping, Fashion|
4.5|
+--------------------+-----------------+--------------------+--------------------+----------+-----+--------------------+-
----+
only showing top 5 rows
```

In [17]:

```
#line20
final_data=joined_data.drop("business_id","categories")
final_data.show(5)
#joined_data = joined_data.select("avg(stars)","stars","name","city","state")
```

```
+-----------------+--------------------+----------+-----+-----+
|       avg(stars)|                name|      city|state|stars|
+-----------------+--------------------+----------+-----+-----+
|3.606060606060606|        Nails Studio|    Dedham|   MA|  3.5|
|4.538461538461538|Beacon Hill Shoe ...|    Boston|   MA|  4.5|
|4.333333333333333| Eric Hollander, DDS|    Austin|   TX|  4.5|
|              3.8|Fresh Touch Cleaners|Winchester|   MA|  4.0|
|4.666666666666667|       Donna & Toots|  Portland|   OR|  4.5|
+-----------------+--------------------+----------+-----+-----+
only showing top 5 rows
```

In [18]:

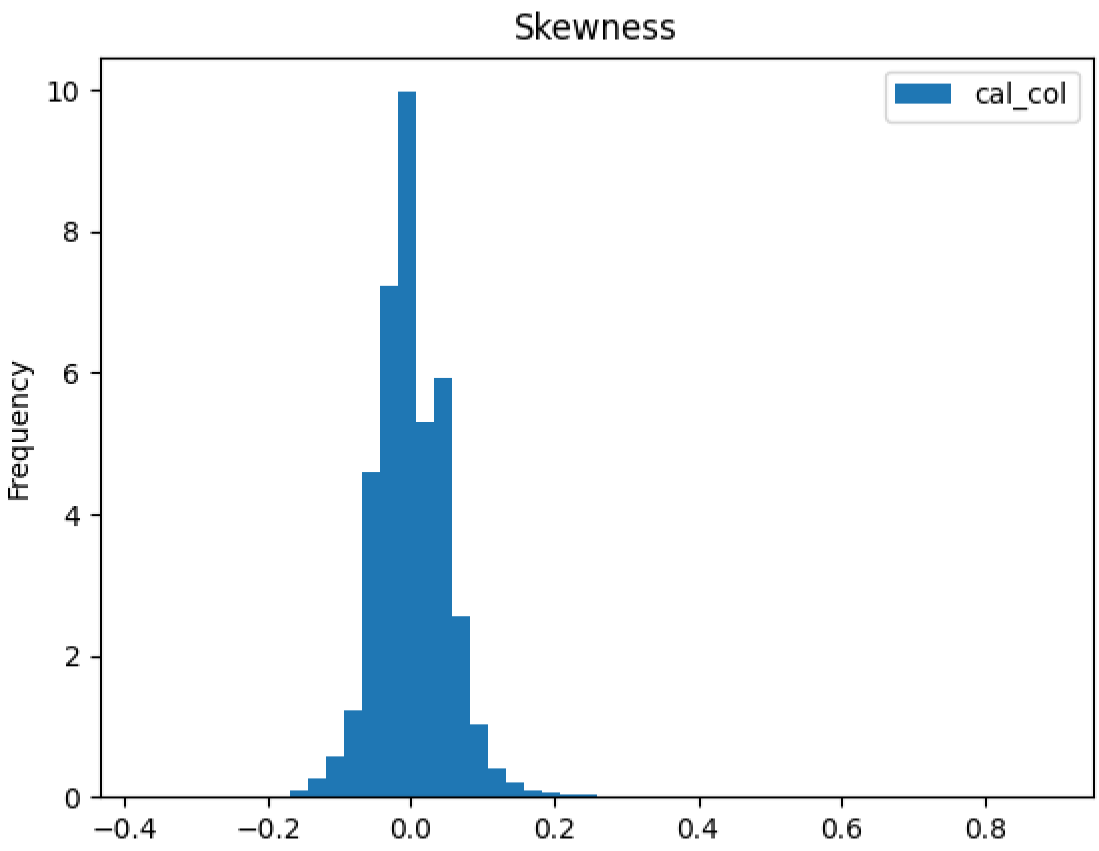
```
#line 21
final_data2 = final_data.withColumn("cal_col",((final_data['avg(stars)'] - final_data['stars']) / final_data['stars'])).c
final_data2.show()
```

```
+--------------------+---------------+-----+--------------------+
|                name|           city|state|             cal_col|
+--------------------+---------------+-----+--------------------+
|        Nails Studio|         Dedham|   MA|0.030303030303030276|
|Beacon Hill Shoe ...|         Boston|   MA|0.008547008547008517|
| Eric Hollander, DDS|         Austin|   TX|-0.03703703703703...|
|Fresh Touch Cleaners|     Winchester|   MA|-0.05000000000000...|
|       Donna & Toots|       Portland|   OR|0.037037037037037104|
|              Subway|       Portland|   OR|-0.01999999999999...|
|Ruiz Branch - Aus...|         Austin|   TX|0.050000000000000044|
|    GibsonBreen & Co|        Atlanta|   GA|                 0.0|
|True Glow | Buckhead|        Atlanta|   GA|0.037037037037037104|
|Brookline Adult &...|      Brookline|   MA|-0.05357142857142857|
|Robert T Franklin...|      Beaverton|   OR|                 0.0|
|             Wendy's|New Westminster|   BC|                 0.0|
|Tarrytown Barber ...|         Austin|   TX|0.047619047619047644|
|     Builder By Bike|       Portland|   OR|                 0.0|
|College Housing N...|       Portland|   OR|-0.04000000000000...|
|Nectar Frozen Yog...|       Portland|   OR|-0.01399176954732...|
|   Searge Sautre, DC|       Chamblee|   GA|-0.02272727272727...|
|        Freedom Tail|         Boston|   MA|                 0.0|
|Reverend Nat's Ha...|       Portland|   OR|-0.02380952380952...|
|    Creekside Kayaks|      Vancouver|   BC|                 0.0|
+--------------------+---------------+-----+--------------------+
only showing top 20 rows
```

In [19]:

```
#line22   #看起来slightly right skewed

final_data2.toPandas().plot(kind="hist",bins=50,title="Skewness",density=True,linestyle = "dashed",linewidth = 1)
%matplot plt
```



In [20]:

```
user_data_ = user.select("user_id","fans","useful")
user_data_.show(truncate=False)
```

```
+----------------------+----+------+
|user_id               |fans|useful|
+----------------------+----+------+
|q_QQ5kBBwlCcbL1s4NVK3g|1357|15038 |
|dIIKEfOgo0KqUfGQvGikPg|1025|21272 |
|D6ErcUnFALnCQN4b1W_TlA|16  |188   |
|JnPIjvC0cmooNDfsa9BmXg|420 |7234  |
|37Hc8hr3cw0iHLoPzLK6Ow|47  |1577  |
|n-QwITZYrXlKQRiV30MqNg|17  |476   |
|eCJoZqpV1fDKJGAsXmWXqQ|1   |53    |
|cojecOwQJpsYDxnjtgzteQ|4   |136   |
|1jXmzuIFKxTnEnR0pxO0Hg|23  |381   |
|-8QoOIfvwwxJ4sY201WP5A|25  |752   |
|EtofuImujQBSo02xa6ZRtQ|5   |159   |
|cxS6dbjyPgPS1S890u_khA|2   |116   |
|MUzkXfPS9JaMgJ907orz0g|86  |2235  |
|tjwblGkWN9m0vsGaypJ0Vw|44  |1469  |
|m-zIVssiXN4bnDFqMdPtEA|0   |6     |
|fxqvyXlml4400BglsxRG_w|38  |903   |
|9edAbpniyhHFdpAvknQPBg|48  |1532  |
|wURnB9fRNGAli13yBwhENA|6   |235   |
|l4P65LXNBnJqI7oTXdBvGg|9   |419   |
|9RIXlhUb_xEVuc_o0QsT0w|2   |44    |
+----------------------+----+------+
only showing top 20 rows
```

In [21]:

```
review = spark.read.json('s3://9760project2/project2_yelp/yelp_academic_dataset_review.json')
```

In [22]:

```
review_data_ = review.select("user_id","useful").withColumnRenamed("useful","users_liked_post").withColumnRenamed("user_i
review_data_.show(truncate=False)
```

```
+----------------------+----------------+
|new_user_id           |users_liked_post|
+----------------------+----------------+
|ak0TdVmGKo4pwqdJSTLwWw|3               |
|YoVfDbnISlW0f7abNQACIg|1               |
|eC5evKn1TWDyHCyQAwguUw|0               |
|SFQ1jcnGguO0LYWnbbftAA|1               |
|0kA0PAJ8QFMeveQWHFqz2A|0               |
|RNm_RWkcd02Li2mKPRe7Eg|2               |
|Q8c91v7luItVB0cMFF_mRA|0               |
|XGkAG92TQ3MQUKGX9sLUhw|0               |
|LWUnzwK0ILquLLZcHHE1Mw|1               |
|99RsBrARhhx60UnAC4yDoA|0               |
|eLAYHxHUutiXswy-CfeiUw|0               |
|Ngl83gs3n22SzLAsNw2znw|3               |
|hn0ZbitvmlHnF--KJGJ6_A|0               |
|B7YSV6r1ePAXc69FkDDuZw|0               |
|xpxWG7jQXZE6BcSeuIq4PQ|0               |
|HvpNr0ohHCaVLp014CQrdw|0               |
|bUHweiErUJ36WGeNrPmEbA|5               |
|JHXQEayrDHOWGexs0dCviA|0               |
|DECuRZwkUw8ELQZfNGef2Q|0               |
|jySmPCkEkJR3cWJlkEs9cw|5               |
+----------------------+----------------+
only showing top 20 rows
```

In [23]:

```
line_23 = user_data_.join(review_data_,review_data_.new_user_id == user_data_.user_id).drop("new_user_id").drop("fans")
line_23.show(truncate=False)
```

```
+----------------------+------+----------------+
|user_id               |useful|users_liked_post|
+----------------------+------+----------------+
|--1UpCuUDJQbqiuFXkOzaw|14    |1               |
|--3Bk72HakneTyp3DEjecg|11    |0               |
|--3Hl2oAvTPlq-f7KtogJg|14    |0               |
|--3Hl2oAvTPlq-f7KtogJg|14    |0               |
|--3Hl2oAvTPlq-f7KtogJg|14    |1               |
|--3Hl2oAvTPlq-f7KtogJg|14    |1               |
|--5FEgQNB3_7WtjxkCsGqA|4     |1               |
|--5FEgQNB3_7WtjxkCsGqA|4     |0               |
|--5FEgQNB3_7WtjxkCsGqA|4     |0               |
|--5FEgQNB3_7WtjxkCsGqA|4     |1               |
|--5FEgQNB3_7WtjxkCsGqA|4     |0               |
|--5FEgQNB3_7WtjxkCsGqA|4     |1               |
|--DCpT4hVZNRpRx572pkEw|2     |0               |
|--DCpT4hVZNRpRx572pkEw|2     |0               |
|--Hh_cXFJJUqYB2STxz1vw|0     |0               |
|--IpFJ0EzvdepaxP47X5eg|3     |0               |
|--IpFJ0EzvdepaxP47X5eg|3     |2               |
|--IpFJ0EzvdepaxP47X5eg|3     |0               |
|--IpFJ0EzvdepaxP47X5eg|3     |0               |
|--IpFJ0EzvdepaxP47X5eg|3     |1               |
+----------------------+------+----------------+
only showing top 20 rows
```

In [24]:

```
final_data__ = line_23.withColumn("useful_ ",(line_23['useful']/ line_23['users_liked_post'])).drop("useful","users_liked
final_data__.show()
```

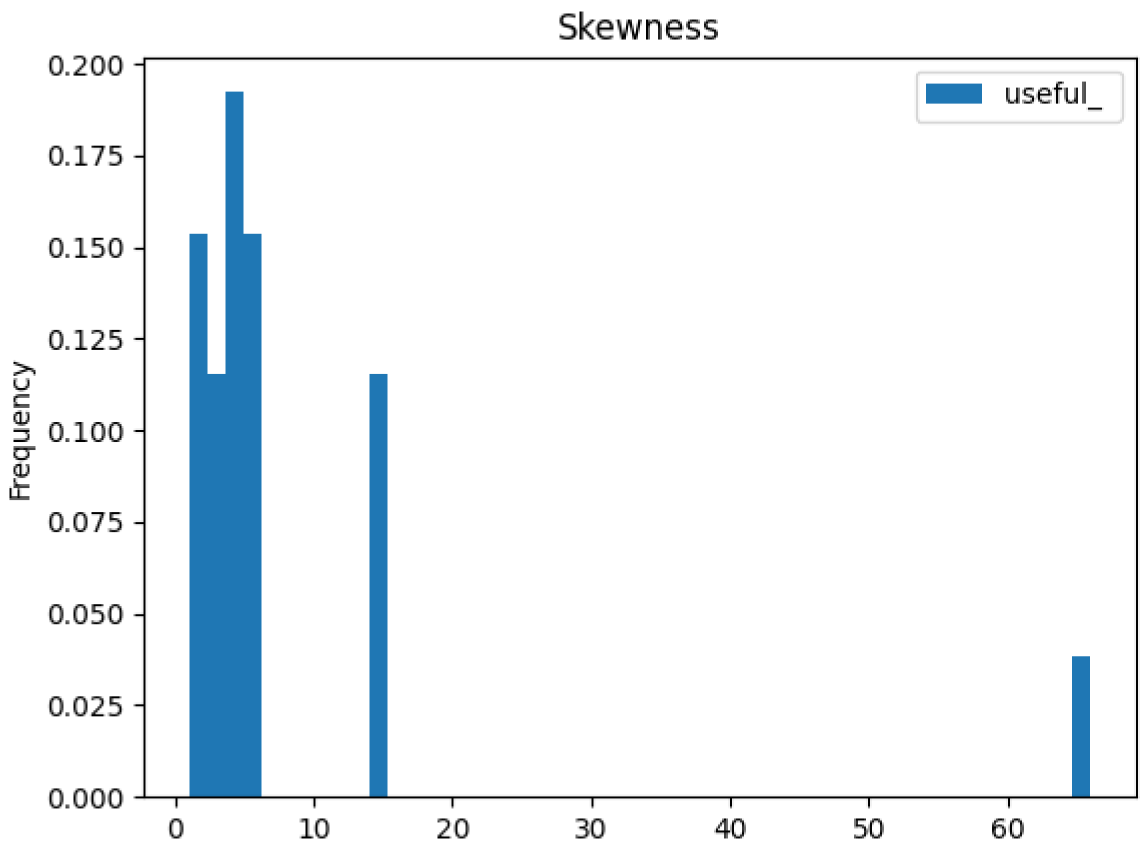
```
+--------------------+------------------+
|             user_id|          useful_ |
+--------------------+------------------+
|--1UpCuUDJQbqiuFX...|              14.0|
|--3Hl2oAvTPlq-f7K...|              14.0|
|--3Hl2oAvTPlq-f7K...|              14.0|
|--5FEgQNB3_7Wtjxk...|               4.0|
|--5FEgQNB3_7Wtjxk...|               4.0|
|--5FEgQNB3_7Wtjxk...|               4.0|
|--IpFJ0EzvdepaxP4...|               1.5|
|--IpFJ0EzvdepaxP4...|               3.0|
|-0-cCufup-5zSCtC9...|               1.0|
|-03gNm8GRPNfgPD4_...|               5.0|
|-03gNm8GRPNfgPD4_...|               5.0|
|-04oKvKUjD6p-wgjU...| 5.651162790697675|
|-08_7TyKsYwY_Jxhg...|               3.0|
|-0CYm85fllm43U7UQ...|               4.0|
|-0PJyCuCFCuUk7_TS...|1.3333333333333333|
|-0PrUwCtOxcoqbvxN...|               2.5|
|-0PrUwCtOxcoqbvxN...|1.6666666666666667|
|-0vUEEyCtW0fE5NL5...|               4.0|
|-12phdDdJ0OpoRVf1...|               6.0|
|-1CV3L7RAk34790wX...|              66.0|
+--------------------+------------------+
only showing top 20 rows
```

In [25]:

```
bar_pandas = final_data__.limit(20).toPandas()
```

In [26]:

```
bar_pandas.plot(kind="hist",bins=50,title="Skewness",density=True,linestyle = "dashed",linewidth = 1)
%matplot plt
```


Skewness
useful_
Frequency
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
50
60

In [ ]: