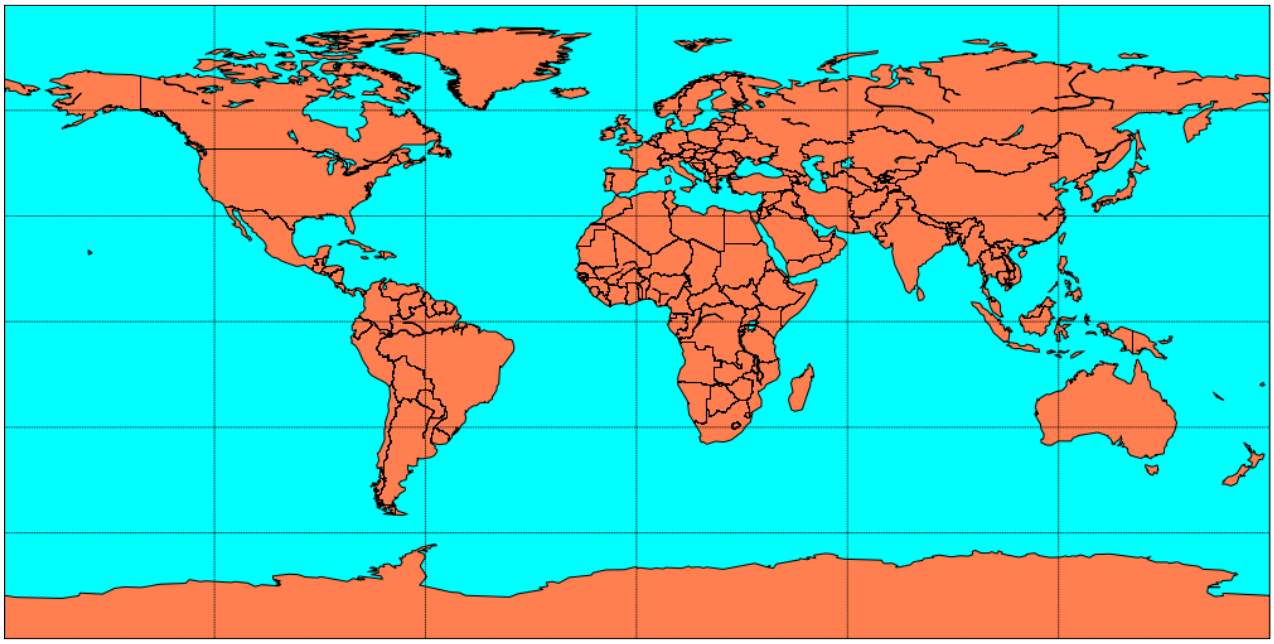
Homework 5

Pavlyuk Lyuba

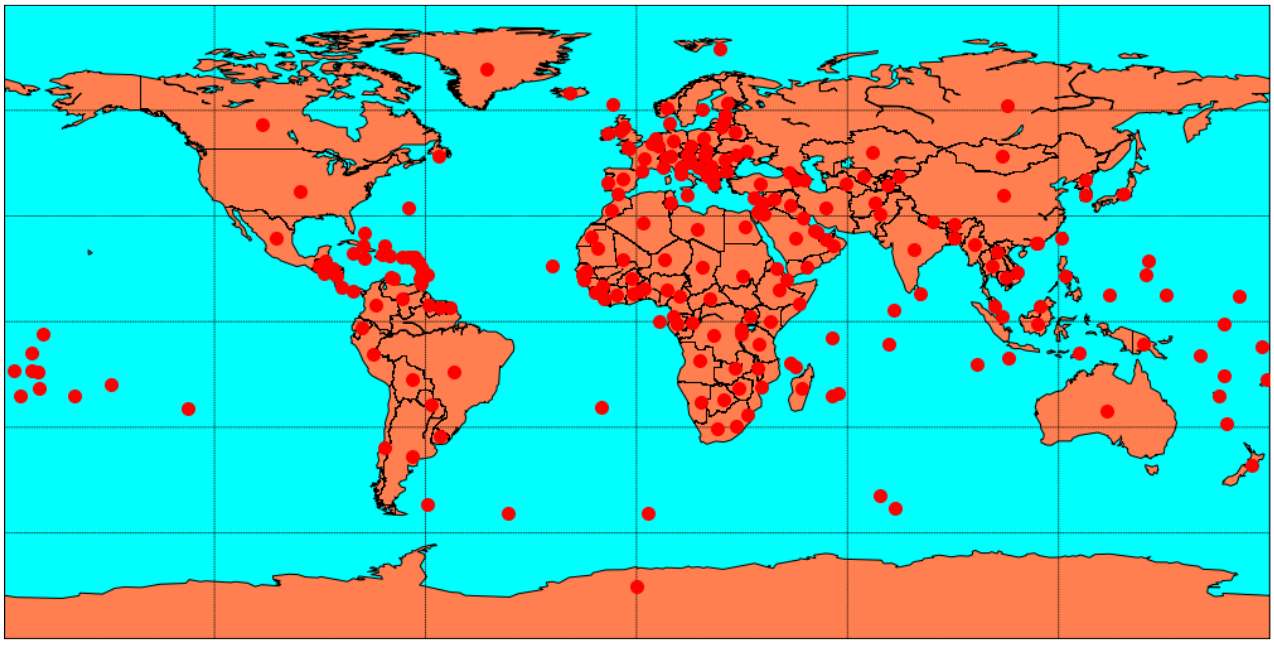
***Задача 1***

**1.** Загрузила файл с данными опроса. Прочитала данные из файла внутри полученного архива survey_results_public.csv с ответами и survey_results_schema.csv с вопросами. Всего 154 вопроса было в опросе и 51392 разработчиков приняло участие в нем.

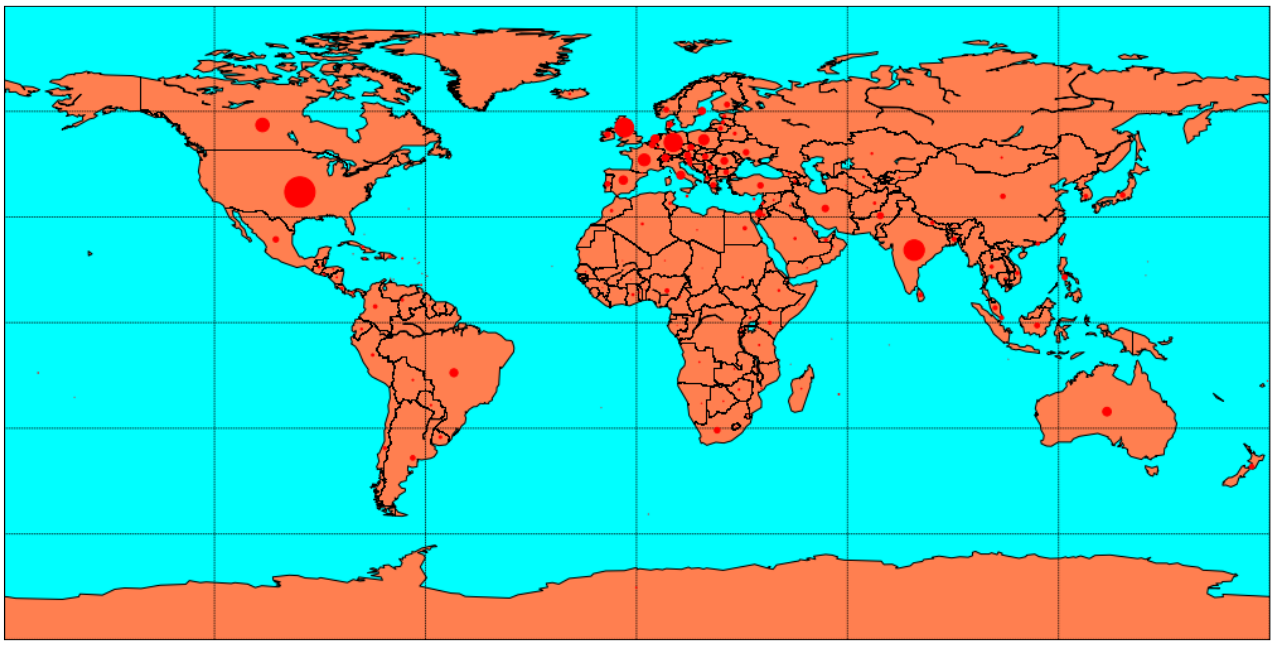
**2.** Подключила библиотеку basemap.
Нарисовала карту мира с границами стран в цилиндрической проекции:



**3.** Скачала координаты центра, загрузила в pandas.DataFrame и нарисовала на карте точку в центре каждой страны:



**4.** Объединила данные по координатам центров стран с данными опроса и вывела в центре каждой страны, из которой был хотя бы один респондент. Круг с площадью пропорционален числу участников опроса для данной страны:



***Задача 2***

**1.** Загрузила файл с данными по грузоперевозкам railways.zip. Прочитала данные из файла внутри полученного архива:

```
['railways201208.csv', 'stations.csv']
```

В нем два файла: railways201808.csv и stations.csv. Загрузила эти два файла в два pandas.DataFrame()

```
railways.head()
```

| | date_priem | fr_code | sto_code | stn_code | dist | weight | amount | taxsum | commodity |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 31.08.2012 | 1100 | 61400 | 3010 | 1944 | 202150 | 139005.0 | 25020.9000 | 8 |
| 1 | 01.08.2012 | 1100 | 61360 | 59250 | 431 | 265500 | 66608.0 | 11989.4400 | 8 |
| 2 | 29.08.2012 | 1100 | 81530 | 60530 | 1443 | 134700 | 74336.0 | 13380.4800 | 8 |
| 3 | 03.08.2012 | 1100 | 62630 | 1030 | 2150 | 67500 | 51431.0 | 9257.5801 | 8 |
| 4 | 02.08.2012 | 1100 | 62710 | 53850 | 1402 | 579150 | 326169.0 | 58710.4220 | 8 |

```
stations.head()
```

| | stshortname | stname | stcode | stdate1 | stdate2 | stroadname | stroadcode | stcountry | stcountrycode |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ВЫБОРГ-ПЕРЕВ | ВЫБОРГ-ПЕРЕВАЛКА | 2340 | 18.09.2000 | 01.01.3000 | ОКТЯБРЬСКАЯ | 1 | Российская Федерация | 643 |
| 1 | КАЛАШНИКОВО | КАЛАШНИКОВО | 6230 | 01.08.2000 | 01.01.3000 | ОКТЯБРЬСКАЯ | 1 | Российская Федерация | 643 |
| 2 | ДОБЫВАЛОВО | ДОБЫВАЛОВО | 5510 | 01.08.2000 | 01.01.3000 | ОКТЯБРЬСКАЯ | 1 | Российская Федерация | 643 |
| 3 | СРЕДНЕРОГАТС | СРЕДНЕРОГАТСКАЯ | 3490 | 16.10.2000 | 01.01.3000 | ОКТЯБРЬСКАЯ | 1 | Российская Федерация | 643 |
| 4 | ЛЕВАШОВО | ЛЕВАШОВО | 3880 | 01.08.2000 | 01.01.3000 | ОКТЯБРЬСКАЯ | 1 | Российская Федерация | 643 |

**2.** Изобразила гистограмму и ядерную оценку плотности для расстояния перевозки и его логарифма:

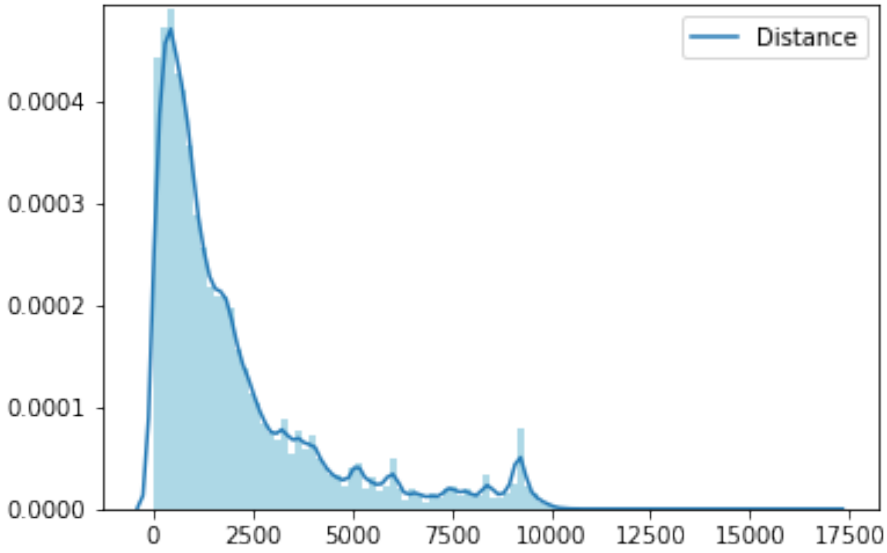



Рис. 1 Гистограмма и ядерная оценка плотности для расстояния перевозки

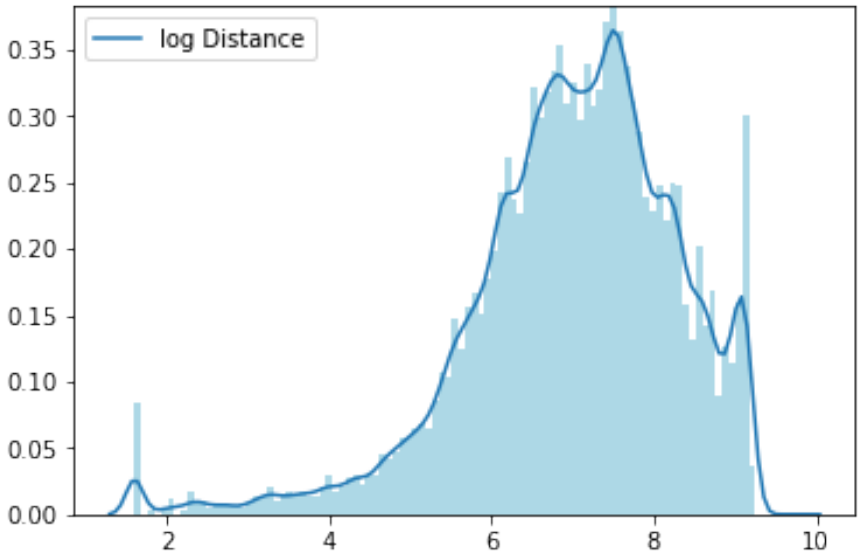



Рис. 2 Гистограмма и ядерная оценка плотности для логарифма расстояния перевозки

**3.** Изобразила логарифм расстояния в виде boxplot, категоризованный по типам грузов (поле commodity) с помощью создания словаря с расшифровкой названий для каждого номера, а затем замены названий в поле commodity. В среднем на самые короткие дистанции перевозят уголь и металлические руды, а на самые длинные зерно и продукты перемола и остальные грузы. Интерквартильный размах распределения лог

дистанции (то есть соответственно 25% (Q1) и 75% (Q3) перцентили) наибольший у перевозок угля, наименьший у удобрений и сырой нефти. Больше всего выбросов по расстоянию (outliers) у металлов, продуктов нефтепереработки, строительных материалов и других грузов, меньше всего у угля и металлических руд.

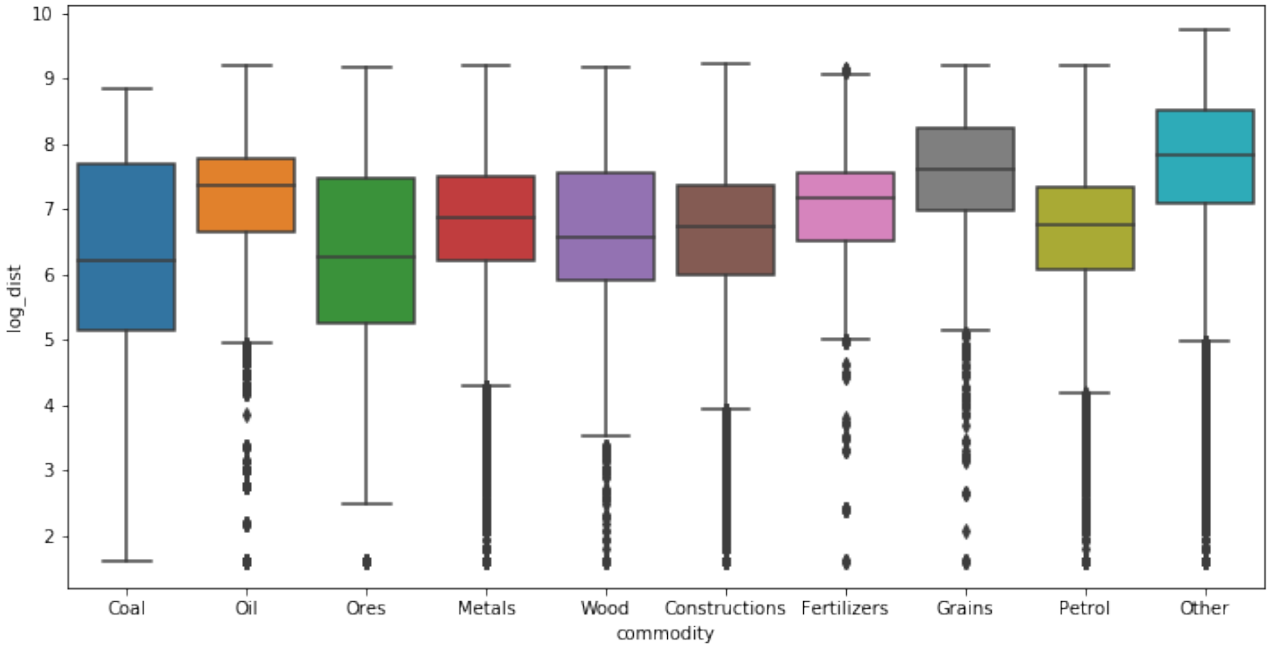



Рис. 3 Логарифм расстояния категоризованный по типам грузов

**4.** Для каждого типа грузов вычислила медиану массы перевозимого груза.

| | commodity | weight |
|---|---|---|
| 0 | Coal | 1.823362e+06 |
| 1 | Constructions | 4.052696e+05 |
| 2 | Fertilizers | 4.055999e+05 |
| 3 | Grains | 1.677138e+05 |
| 4 | Metals | 1.445002e+05 |
| 5 | Oil | 5.448691e+05 |
| 6 | Ores | 3.884998e+06 |
| 7 | Other | 6.286757e+04 |
| 8 | Petrol | 3.012584e+05 |
| 9 | Wood | 1.186295e+05 |

Относительно медианы разбила грузы на тяжелые и легкие по каждой категории. Изобразила violinplot с распределениями расстояния перевозки, классифицированные по типам грузов, в котором слева и справа распределения грузов относительно небольшой и большой массы соответственно.

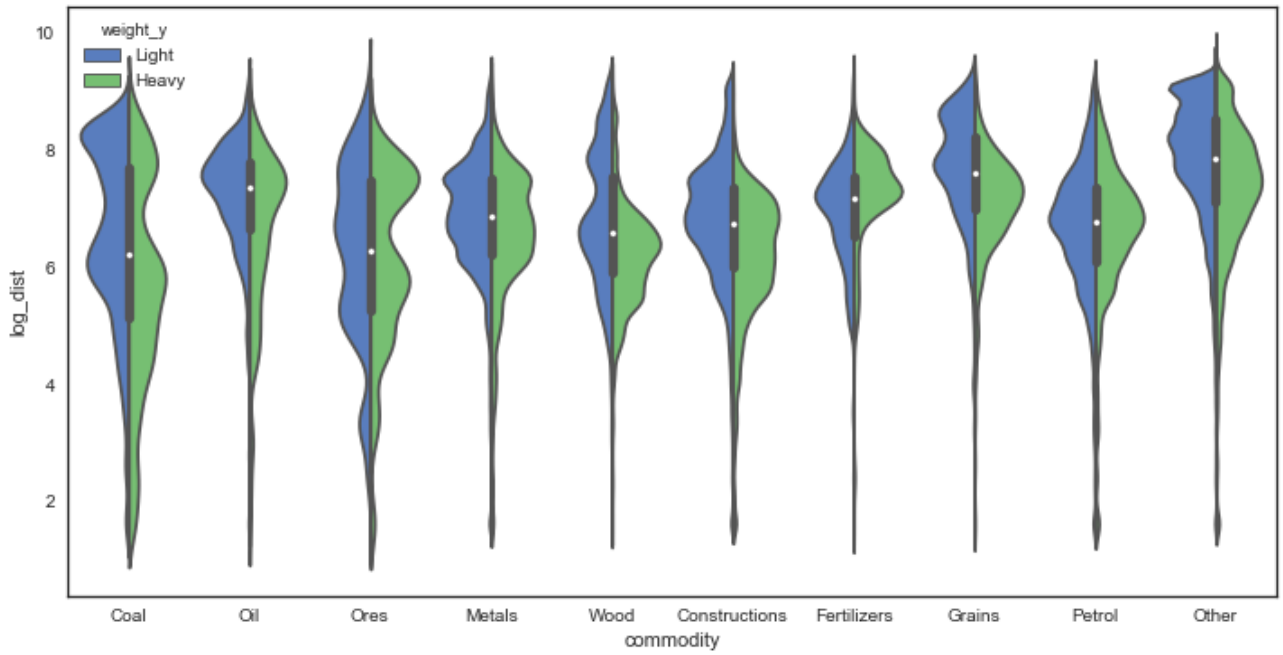



Рис. 4 Распределение расстояния перевозки, классифицированные по типам грузов

**5.** Изобразила диаграмму рассеивания (scatterplot), для которой по горизонтали логарифм произведения расстояния на массу груза, а по вертикали логарифм провозной платы:

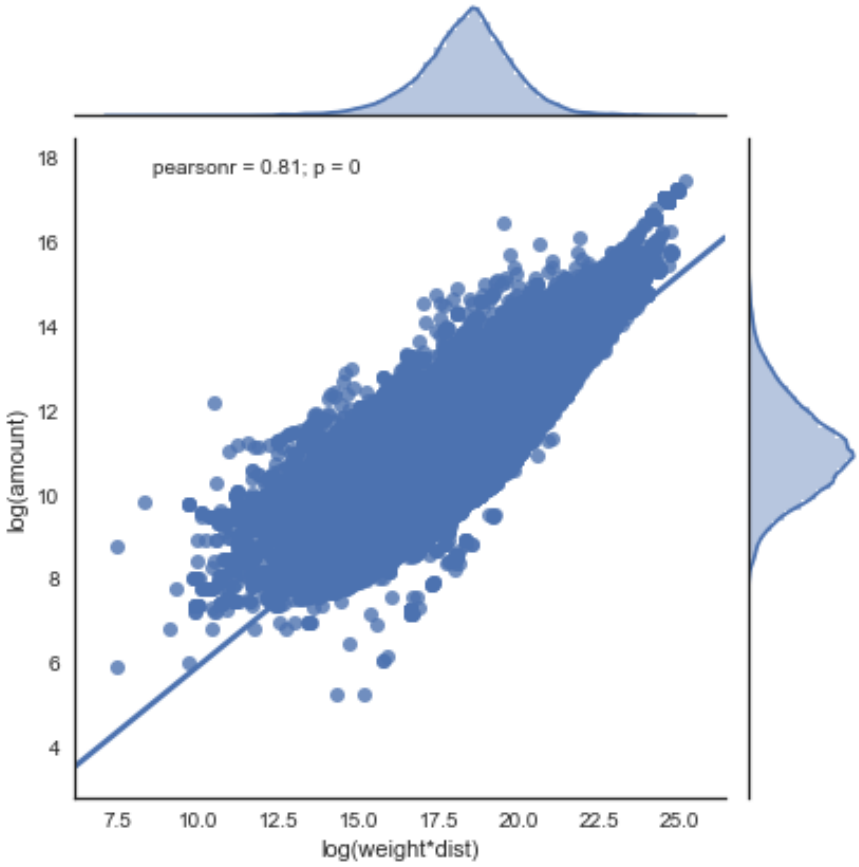



Рис. 5 Диаграмма рассеивания логарифма провозной платы от логарифма произведения расстояния на массу груза

Для того, чтобы подробнее посмотреть на взаимосвязь двух численных признаков использовала joint plot — это гибрид scatter plot и histogram. Посмотрела на то, как связаны между собой логарифм произведения расстояния на массу груза и логарифм провозной платы (зависимость прямо пропорциональная).

Наблюдения, для которых провозная плата равна нулю были выброшены из выборки, для того что не увеличивать ошибку при построении прямой ( логарифма от нуля не существует).