# Отчет по 3-ей лабораторной работе

## Выполнил Михайлов Денис

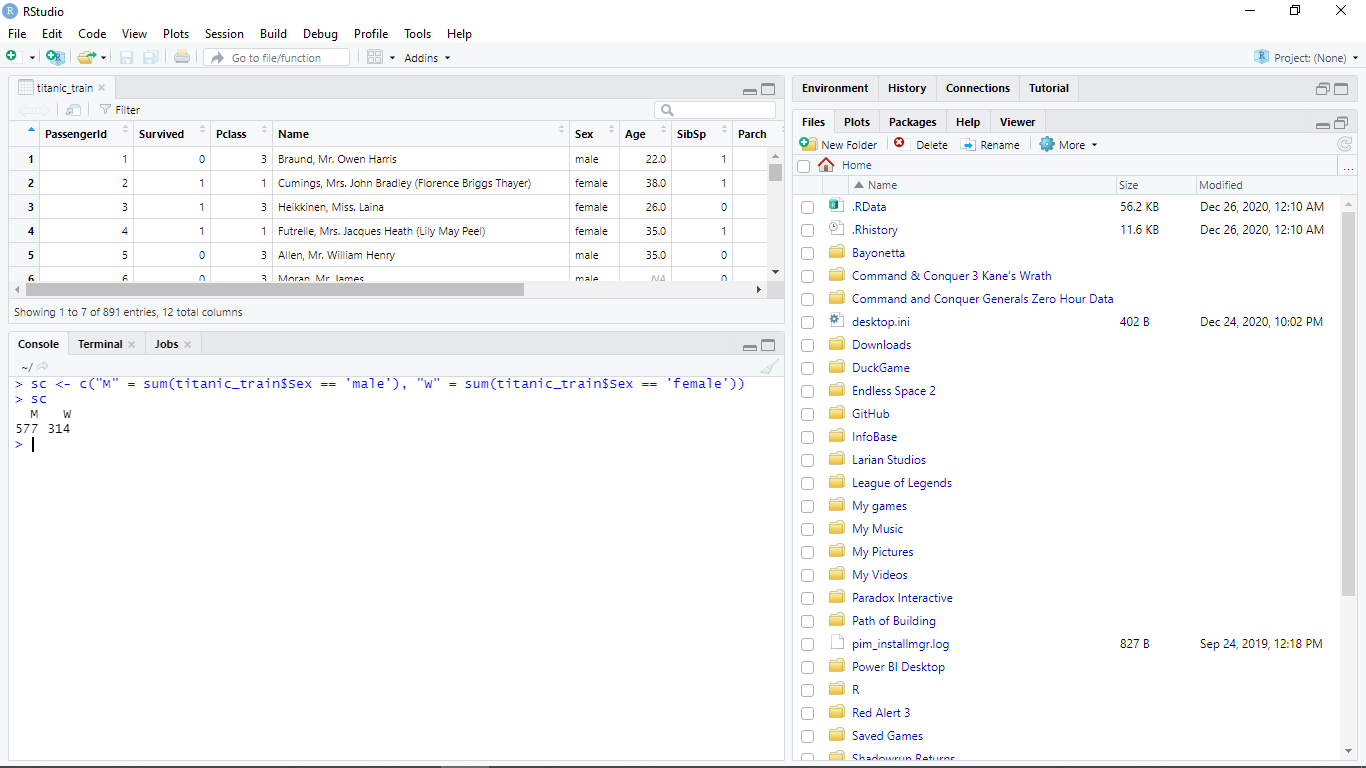
## Студент группы Б8117-02.03.01

### Введение

Данная лабораторная работа выполнялась в среде RStudio. Все операции проводились на базой данных «titanic\_train.csv».

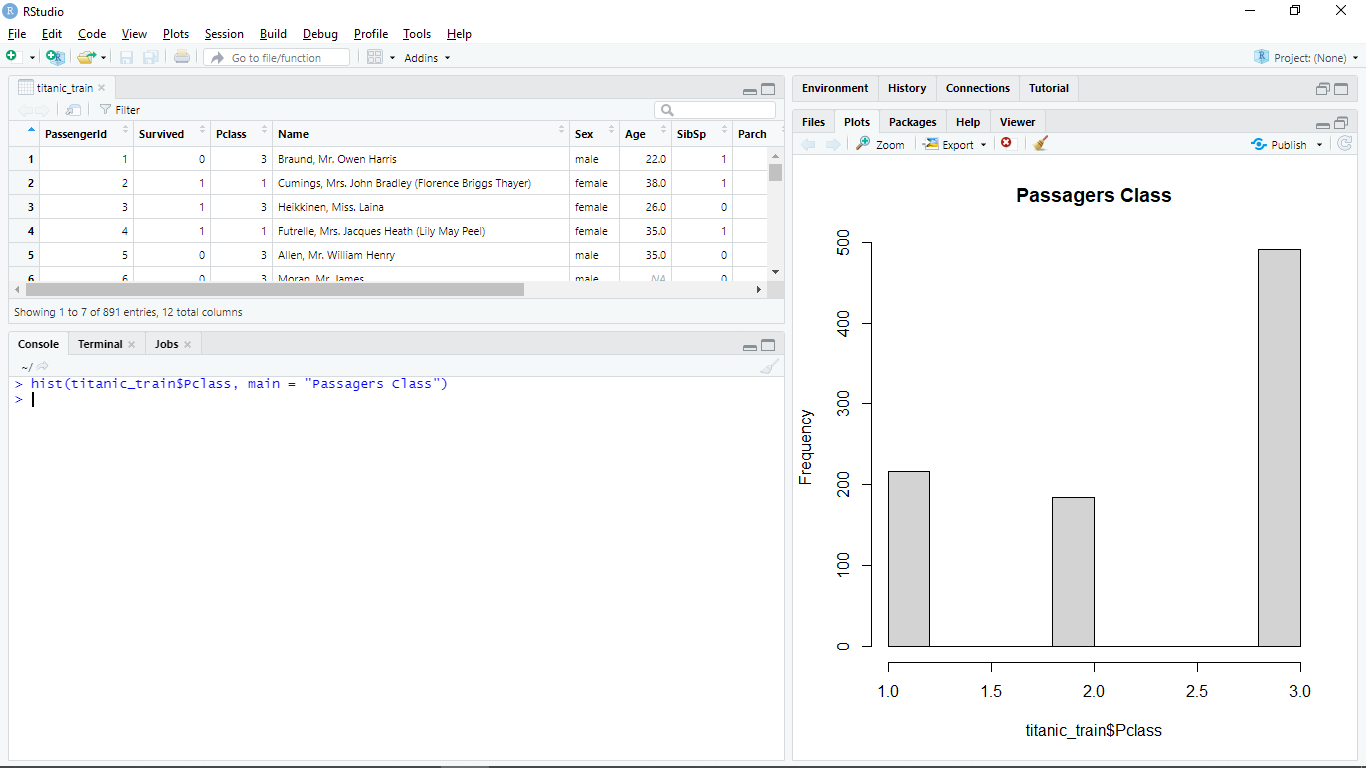
Подключение к БД: *read.csv('titanic\_train.csv')*

### Задание 1

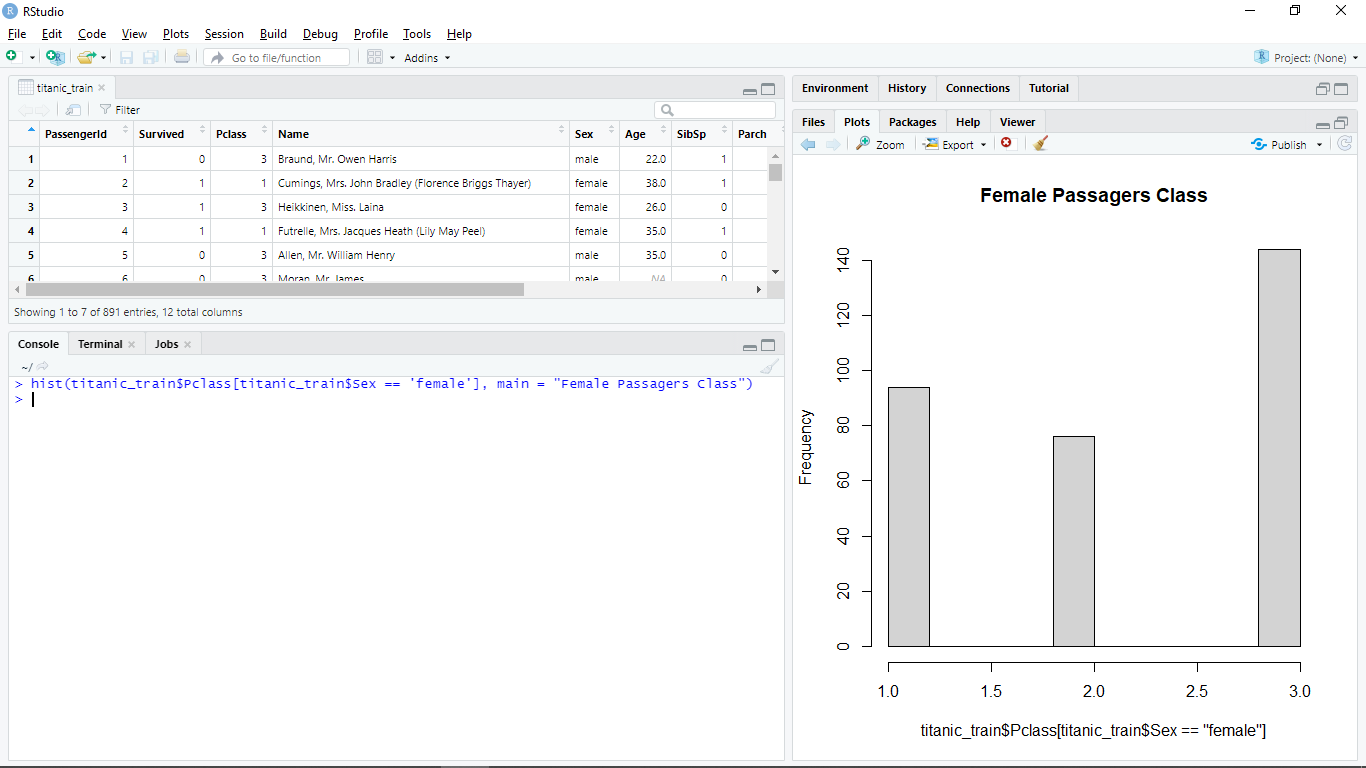
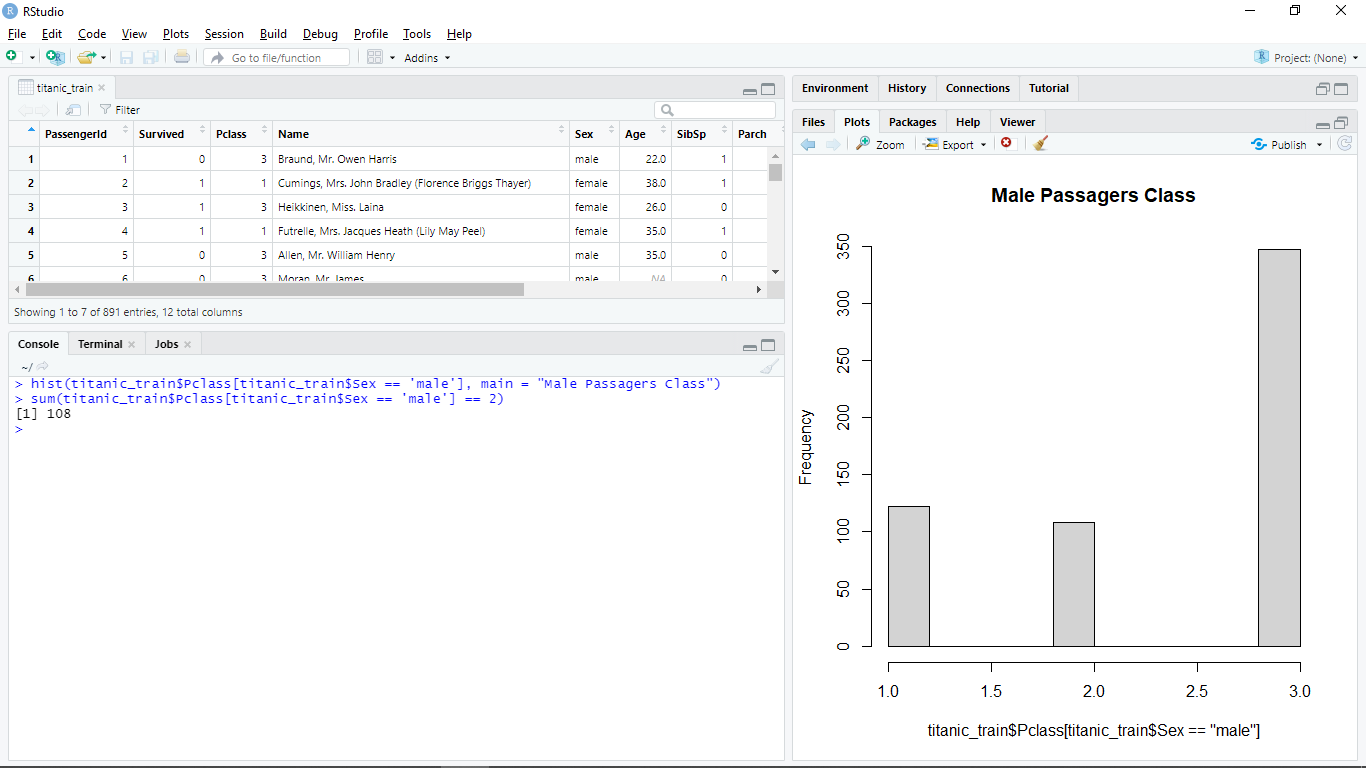
Поставленный вопрос: Сколько мужчин / женщин находилось на борту?

Создаем вектор (функция *с()* ) из двух элементов: количество строк, удовлетворяющие условию *titanic\_train[‘Sex’] == ‘male’,* и количество строк, удовлетворяющие условию *titanic\_train[‘Sex’] == ‘female’*.

### Задание 2

Поставленный вопрос: Выведите распределение переменной Pclass (социально-экономический статус) и это же распределение, только для мужчин / женщин по отдельности. Сколько было мужчин 2-го класса?

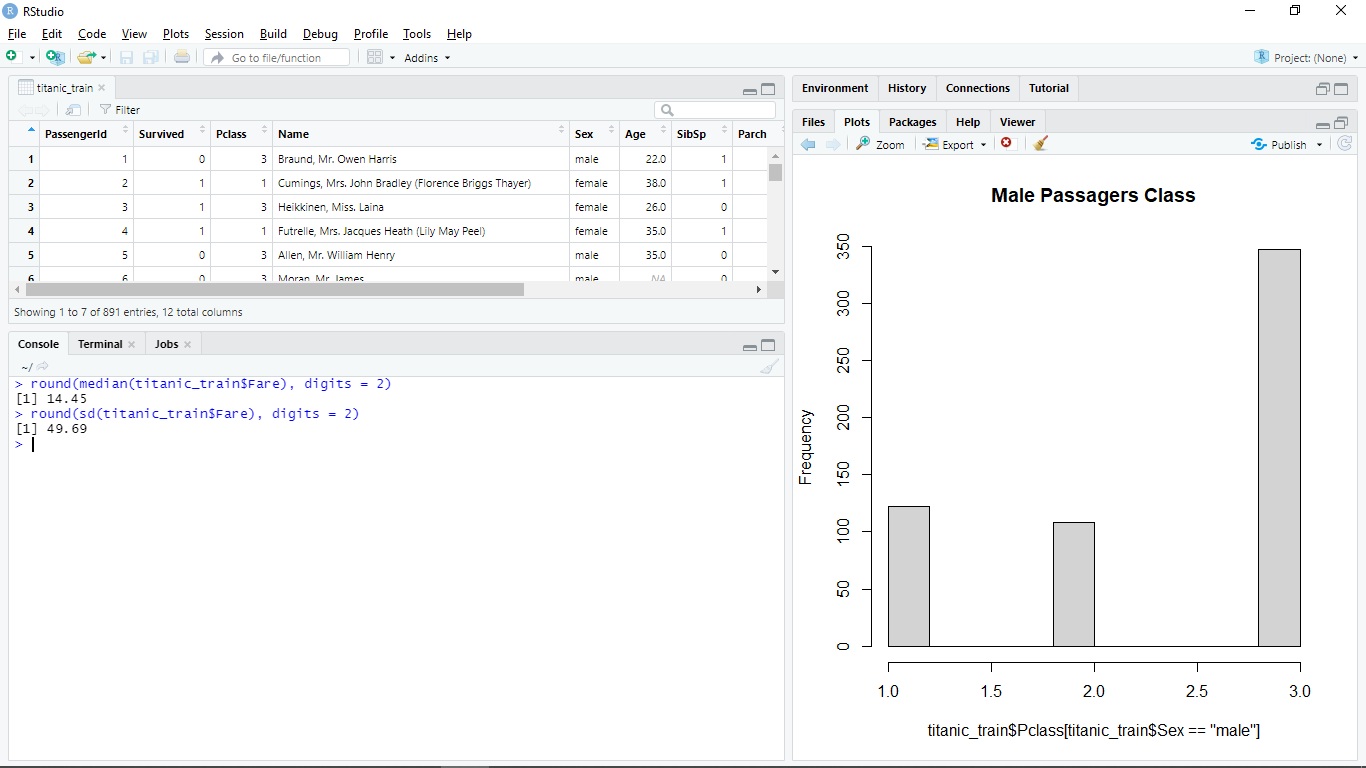
На первом скриншоте представлено распределение переменной PClass c помощью встроенного метода *hist,* который вывод гистограмму по заданным параметрам.

В следующих двух скриншотах будет использована тот же метод. Первый скриншот – распределение переменной PClass только для женщин, второй – только для мужчин.

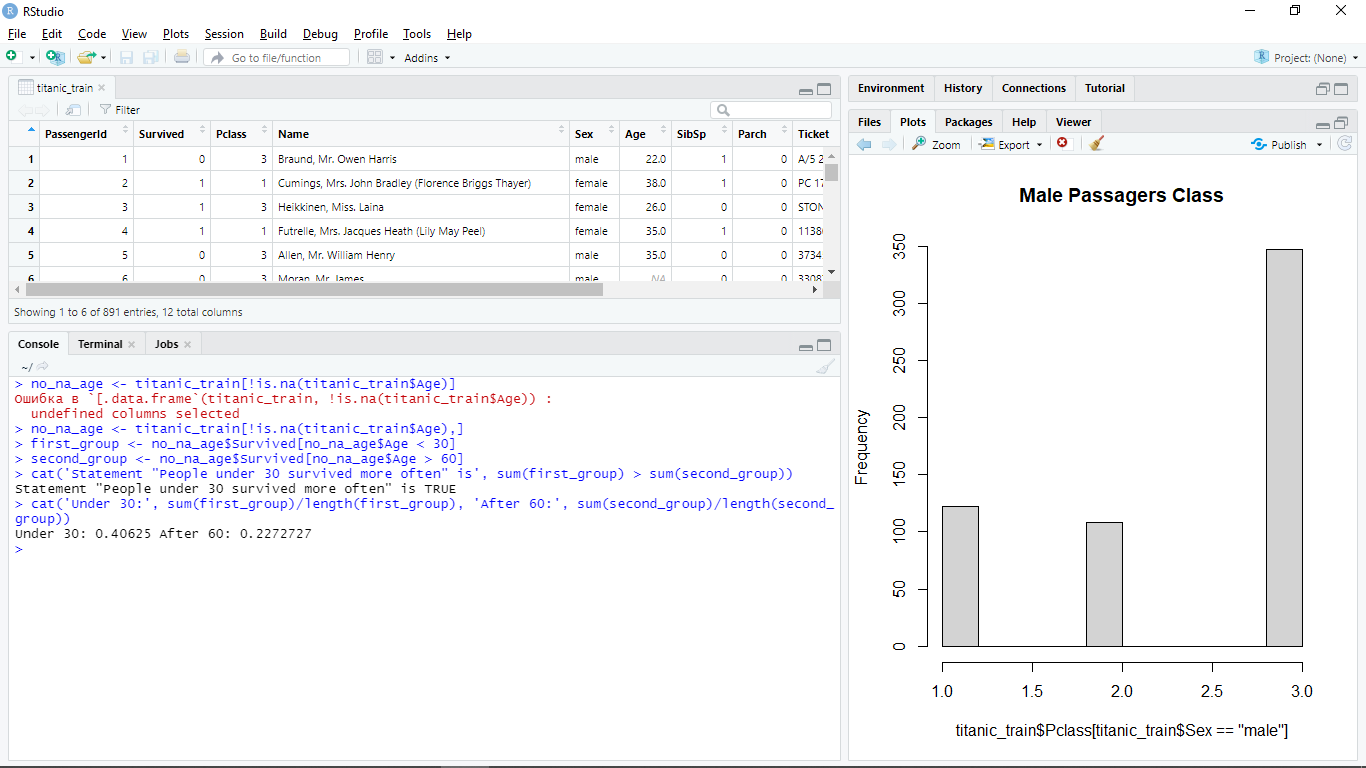
Как видно на последнем скриншоте, вызывается функция *sum* для подсчета количества строк после фильтрации таблицы по столбцам Sex и PClass.

### Задание 3.

Поставленный вопрос: Каковы медиана и стандартное отклонение платежей (Fare)? Округлите до 2 десятичных знаков.

Функция *median*, как следует из названия, посчитывает медиану для столбца значений Fare. Функция *sd* посчитывает стандартное отклонение. Функция *round* округляет число по параметру *digits*. Результат представлен на скриншоте ниже.

### Задание 4.

Поставленный вопрос: Правда ли, что люди моложе 30 лет выживали чаще, чем люди старше 60 лет? Каковы доли выживших в обеих группах?

Переменная no\_na\_age хранит в себе исходную таблицу, отфильтрованную по столбцу Age с помощью метода *is\_na*. Этот метод возвращает логический вектор той же длины, что и входящий вектор, заполняя ее TRUE, если элементы являются NA, и FALSE, если значение не является NA. Так как нам необходима таблица без пустых элементов по столбцу Age, то ставим знак отрицания «!» перед методом.

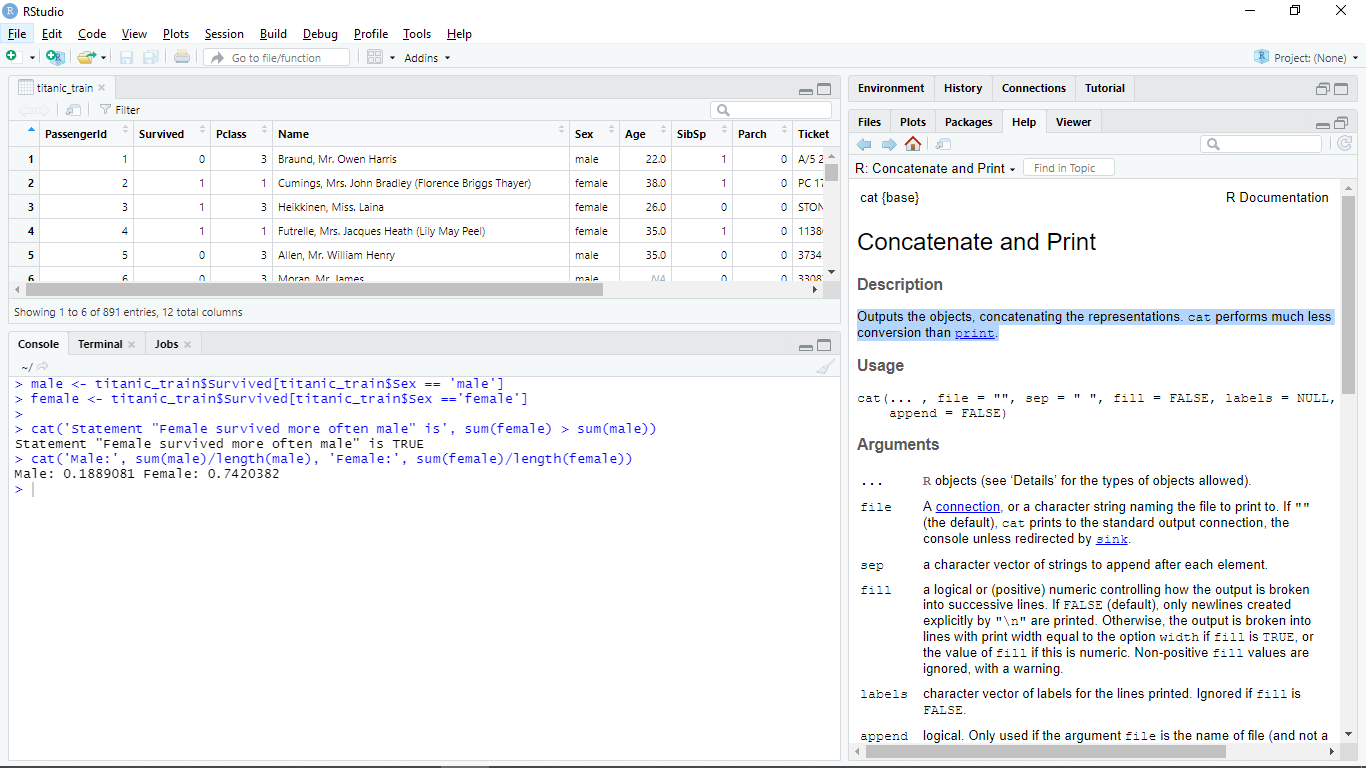
Переменная first\_group хранит в себе людей, чей возраст меньше 30-ти лет. Переменная second\_group хранит в себе людей, чей возраст больше 6-ти лет.

Метод *cat* работает, как и метод *print*, но более упрощенно. Как видно из первого сообщения, люди моложе 30 лет выживали чаще.

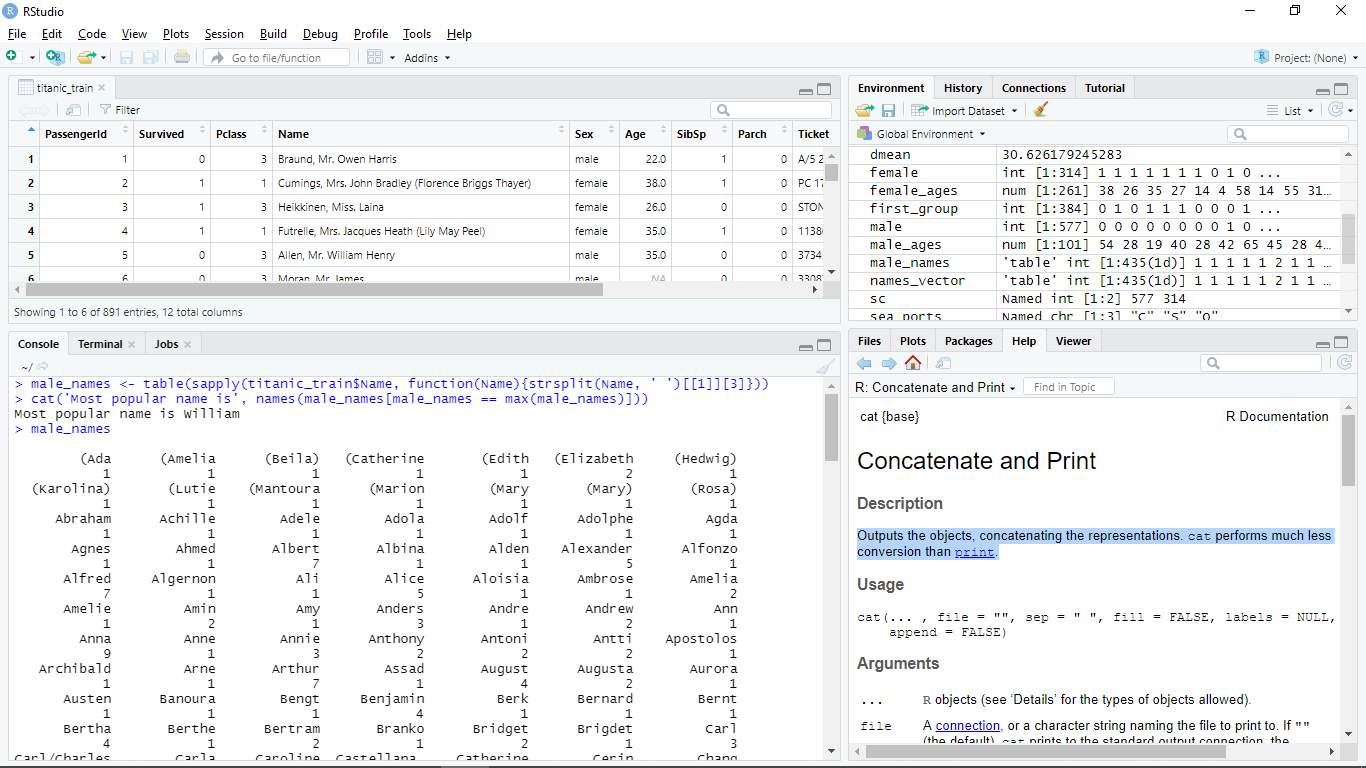
Во втором сообщении видно коэффициент выживших по отношению ко всем пассажирам возрастной группы.

### Задание 5.

Поставленный вопрос: Правда ли, что женщины выживали чаще мужчин? Каковы доли выживших в обеих группах?

Задание аналогично предыдущему, поэтому просто повторяем все действия, кроме создания no\_na\_age. Он не потребуется. Результаты представлены на скриншоте ниже.

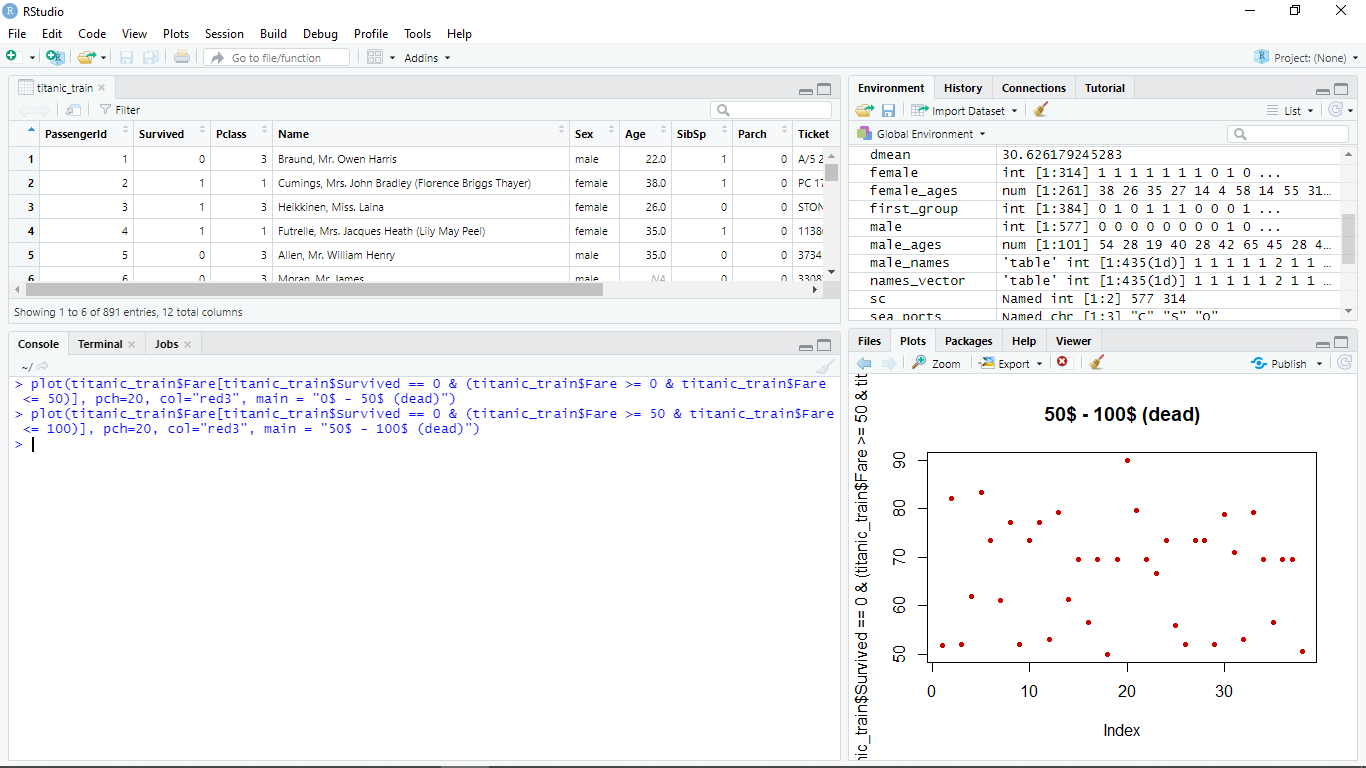
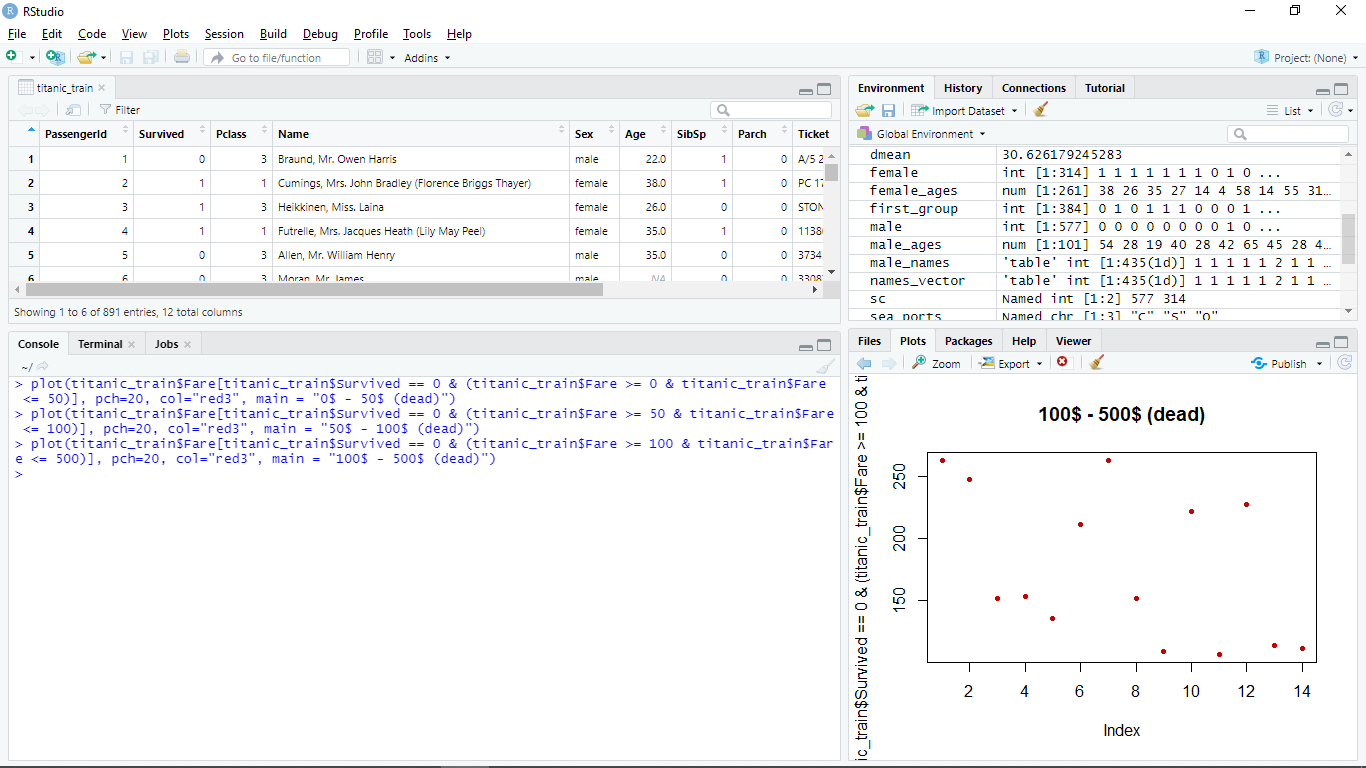
### Задание 6.

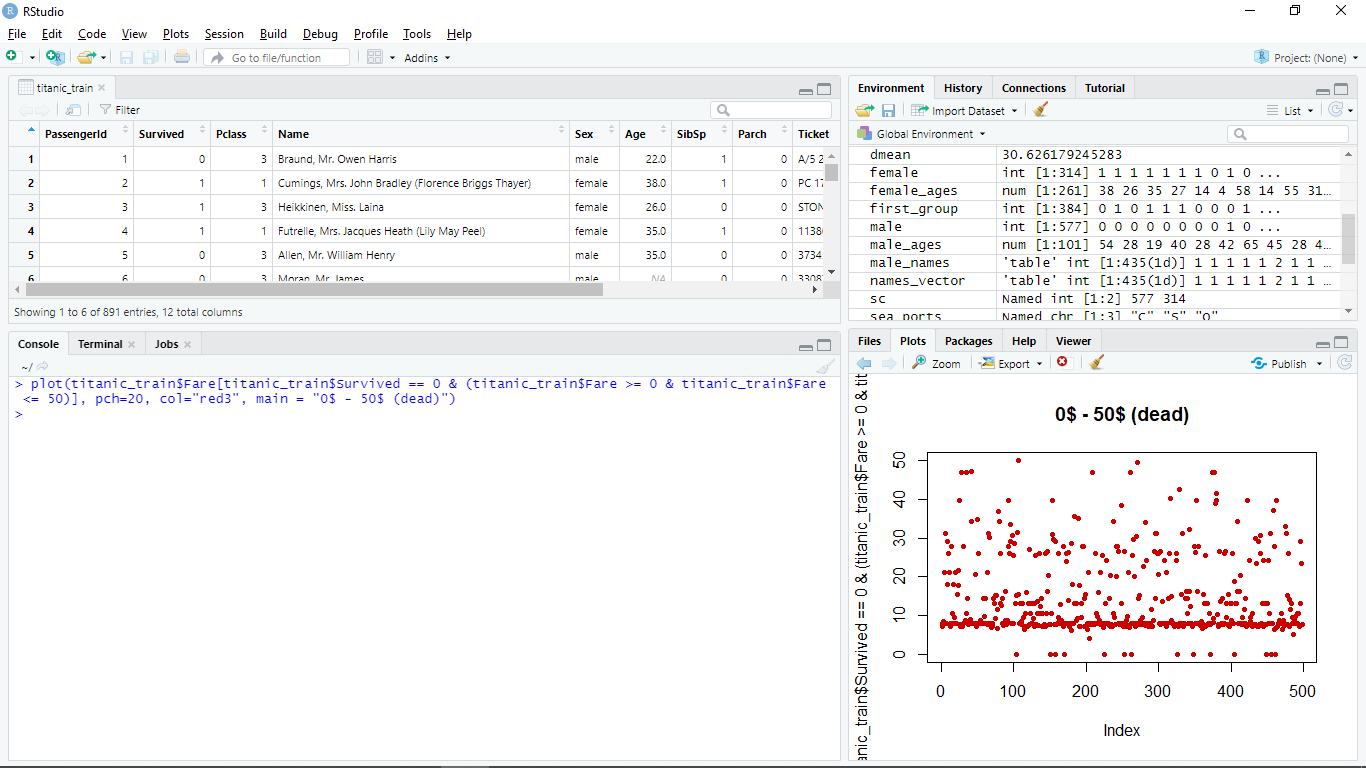
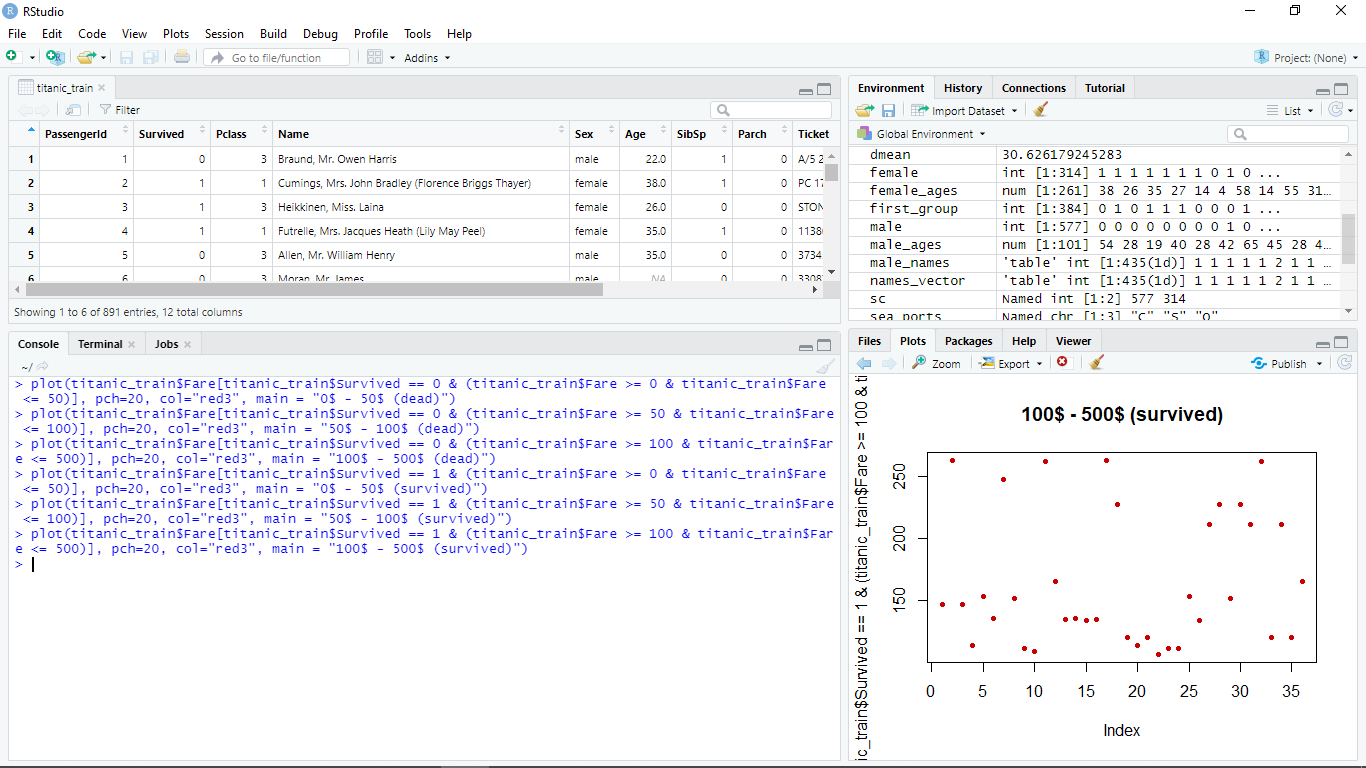
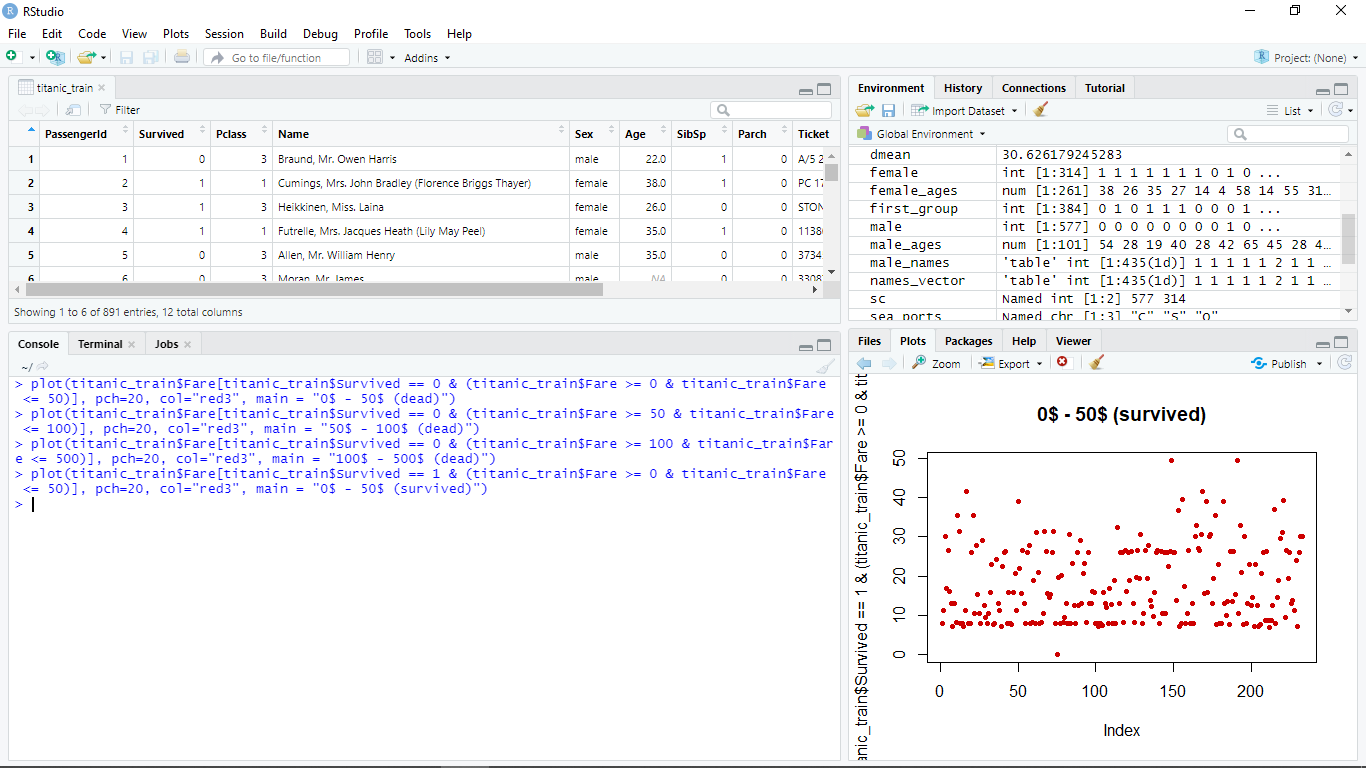
Поставленный вопрос: Найдите самое популярное имя среди пассажиров Титаника мужского пола?

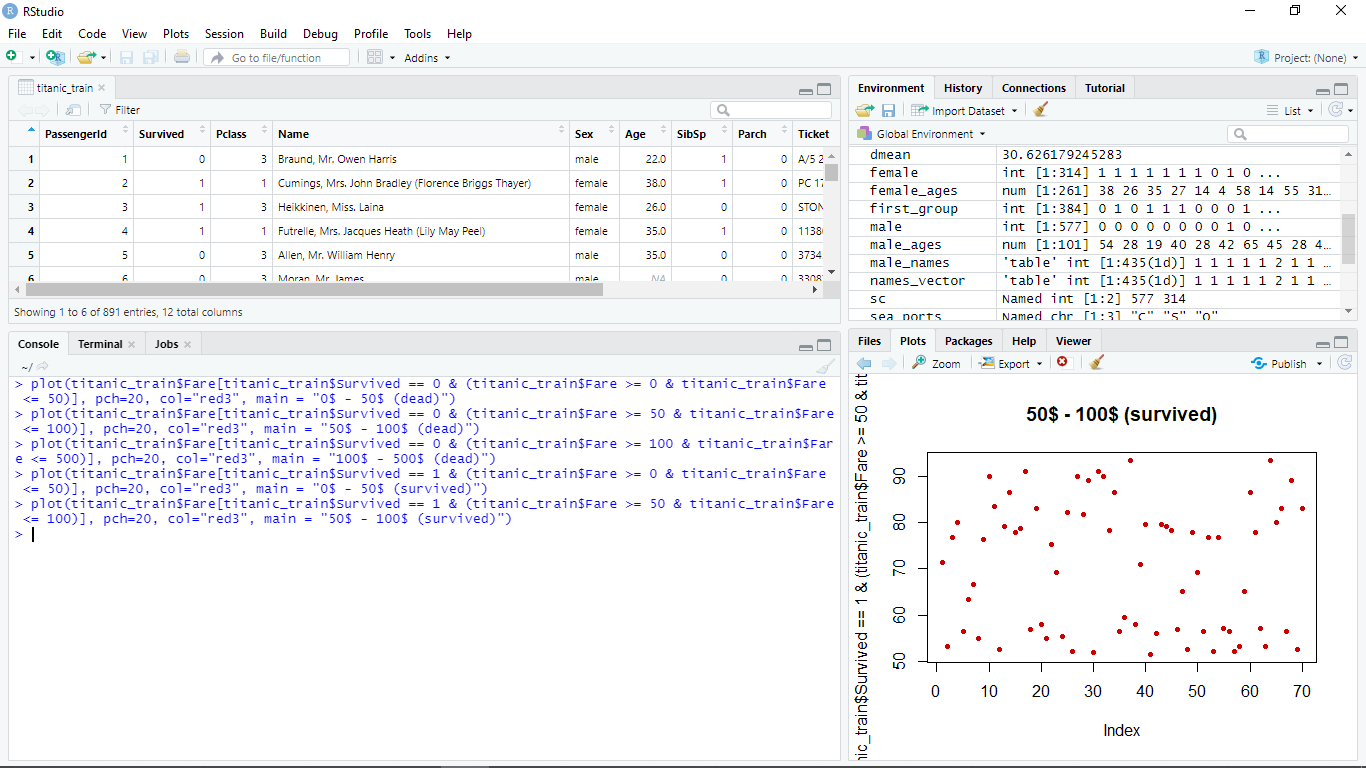
Метод table формирует словарь. В данном случае, он будет формировать словарь из имени пассажира и количества его повторений. Так как имя пассажира состоит из имени и фамилии, с помощью метода sapply формируем вектор, в котором содержатся только имена пассажиров.

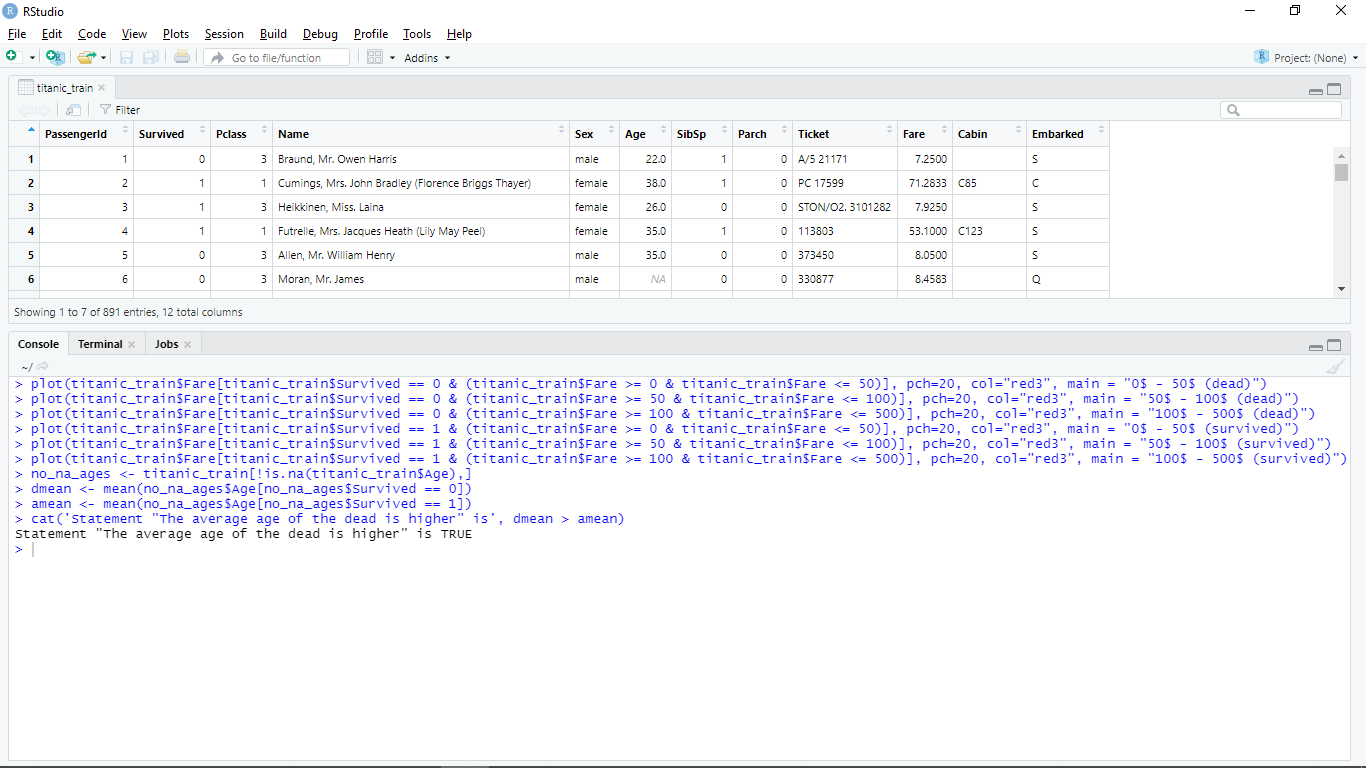
Результат показан на скриншоте.

### Задание 7.

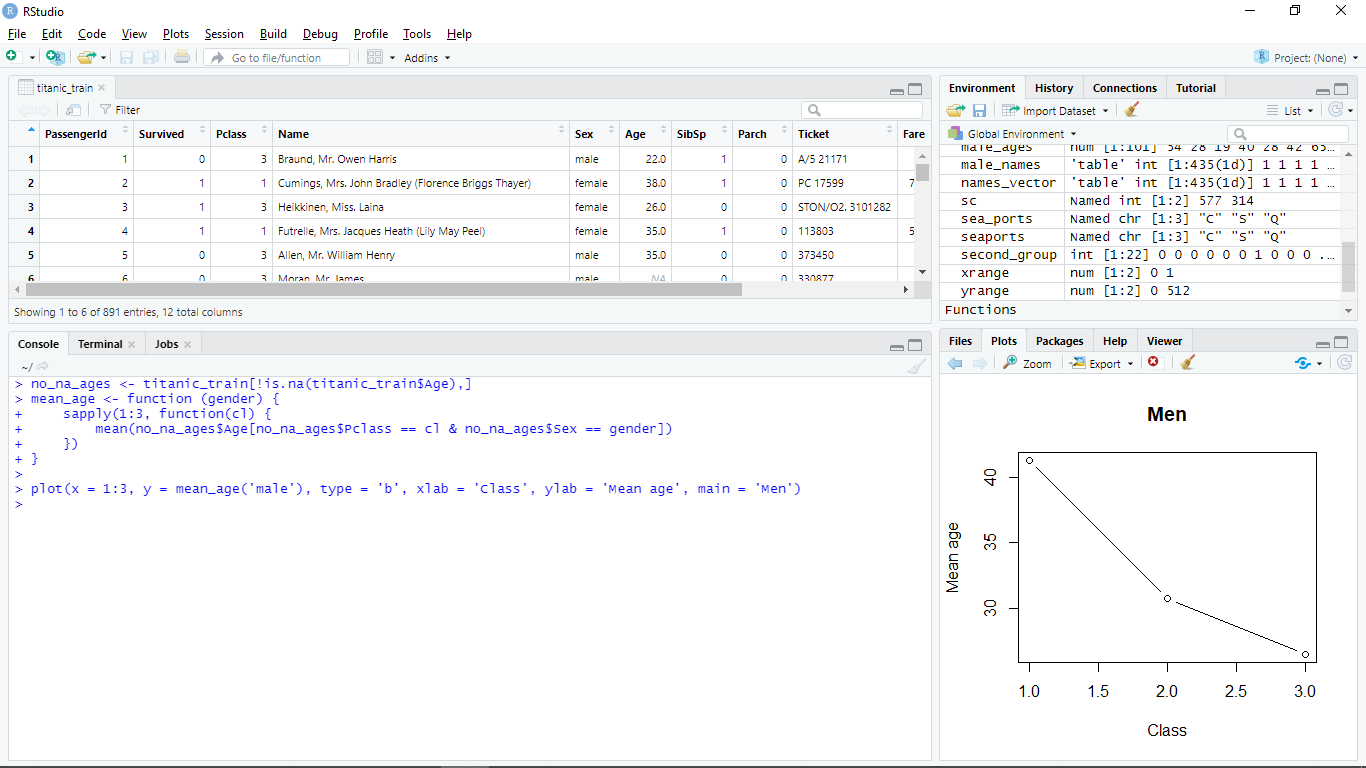
Поставленный вопрос: Сравните графически распределение стоимости билетов у спасенных и у погибших. Средний возраст погибших выше, верно?

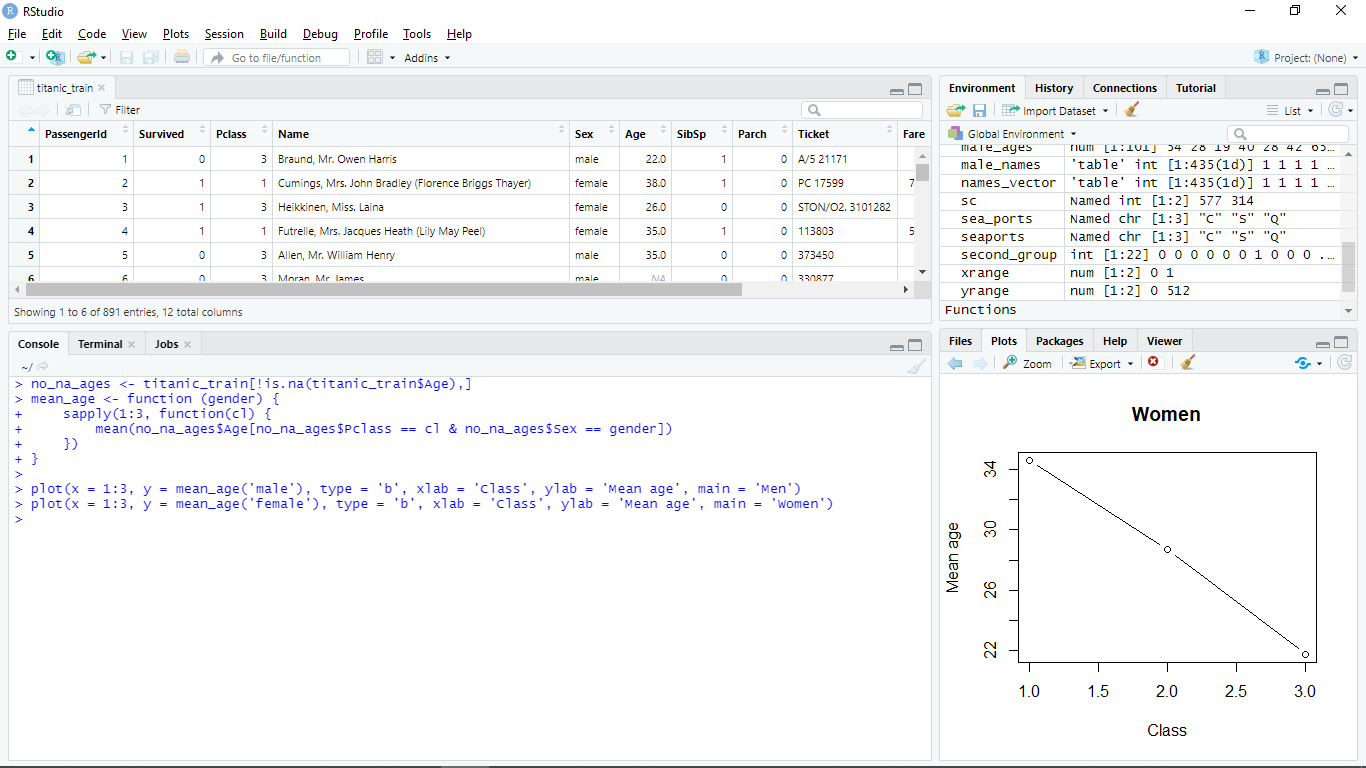
Выше показаны графическое распределение стоимости билетов у погибших.

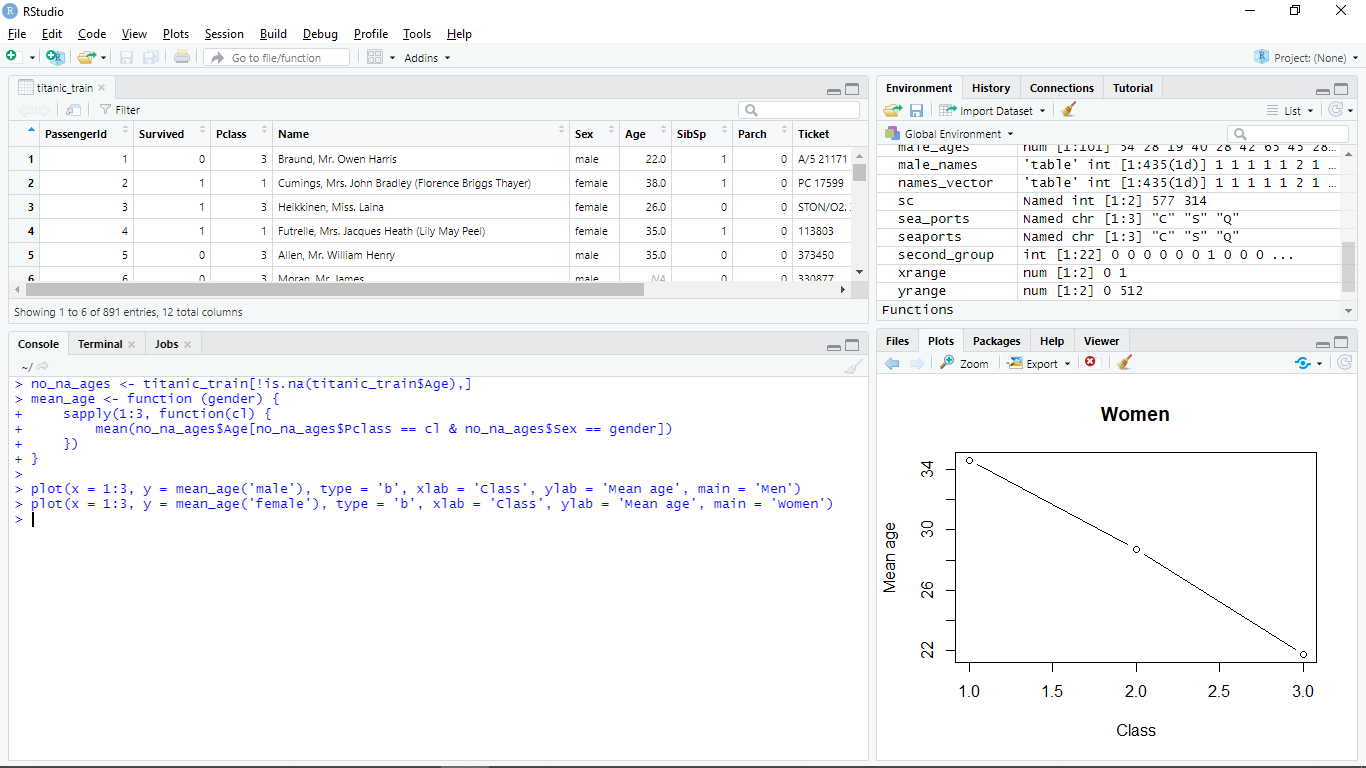
Выше показаны графическое распределение стоимости билетов у выживших.

Далее, требуется посчитать средний возраст погибших и выживших, после чего сравнить их. Функция *mean* вычисляет средние значение. Следующий скриншот показывает результат сравнения, а также все команды, которые были введены.

### Задание 8.

Поставленный вопрос: Как отличается средний возраст мужчин / женщин в зависимости от класса обслуживания?

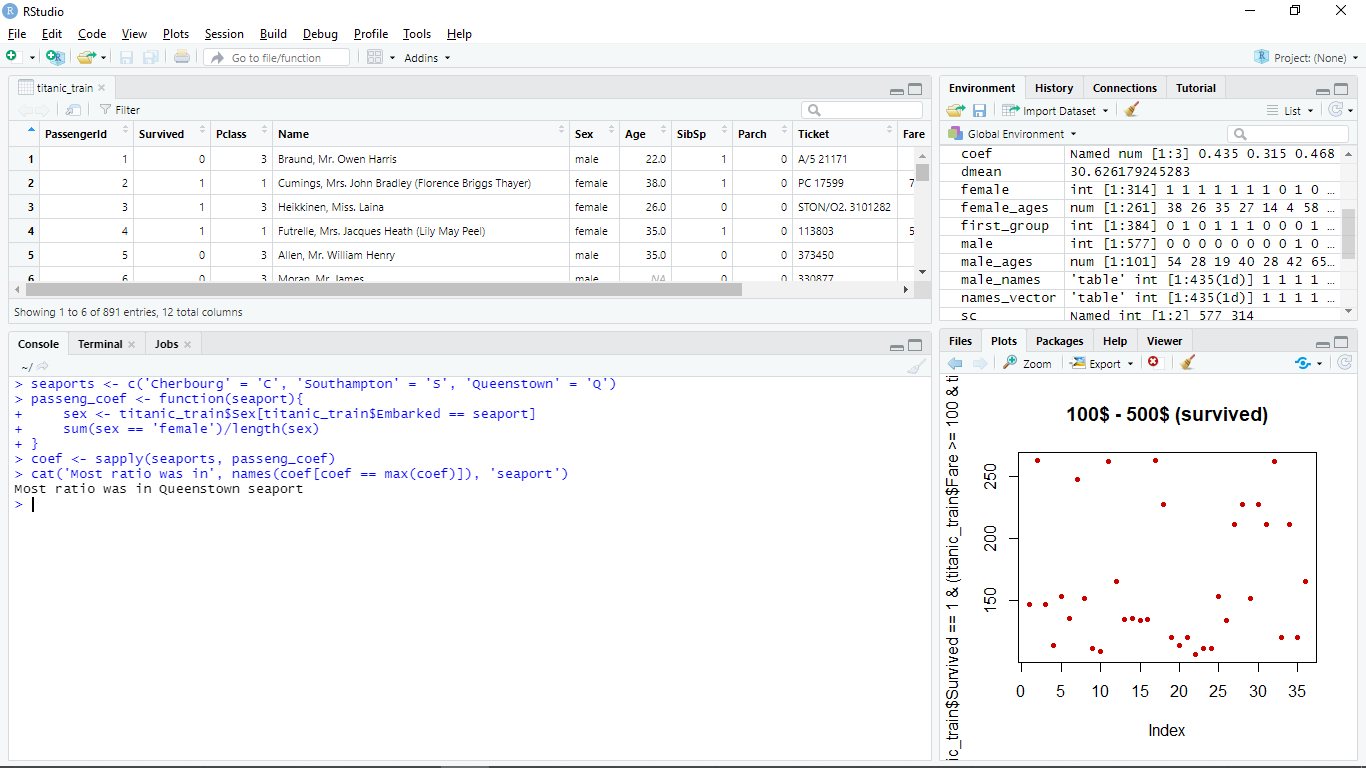
Первый скриншот показывает график зависимости среднего возраста мужчин от класса обслуживания. Второй скриншот - график зависимости среднего возраста мужчин от класса обслуживания.



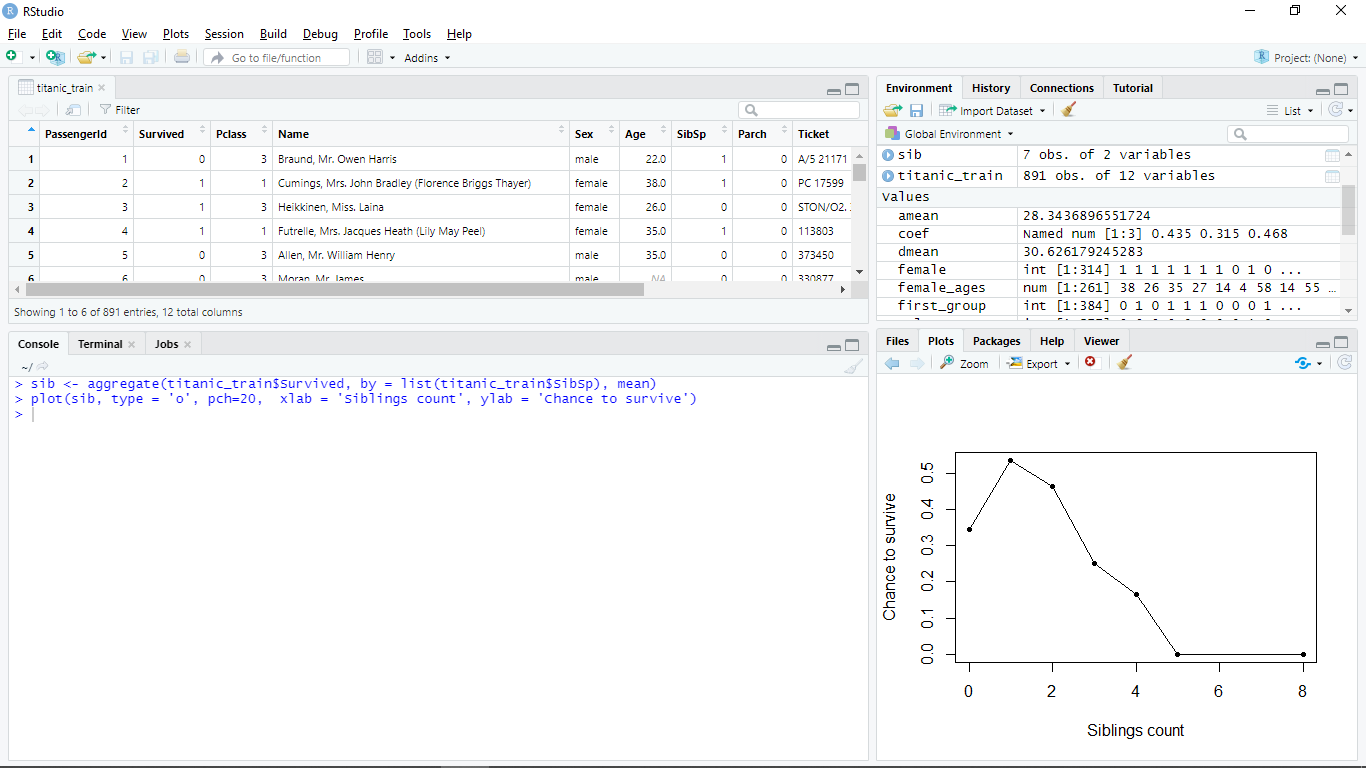
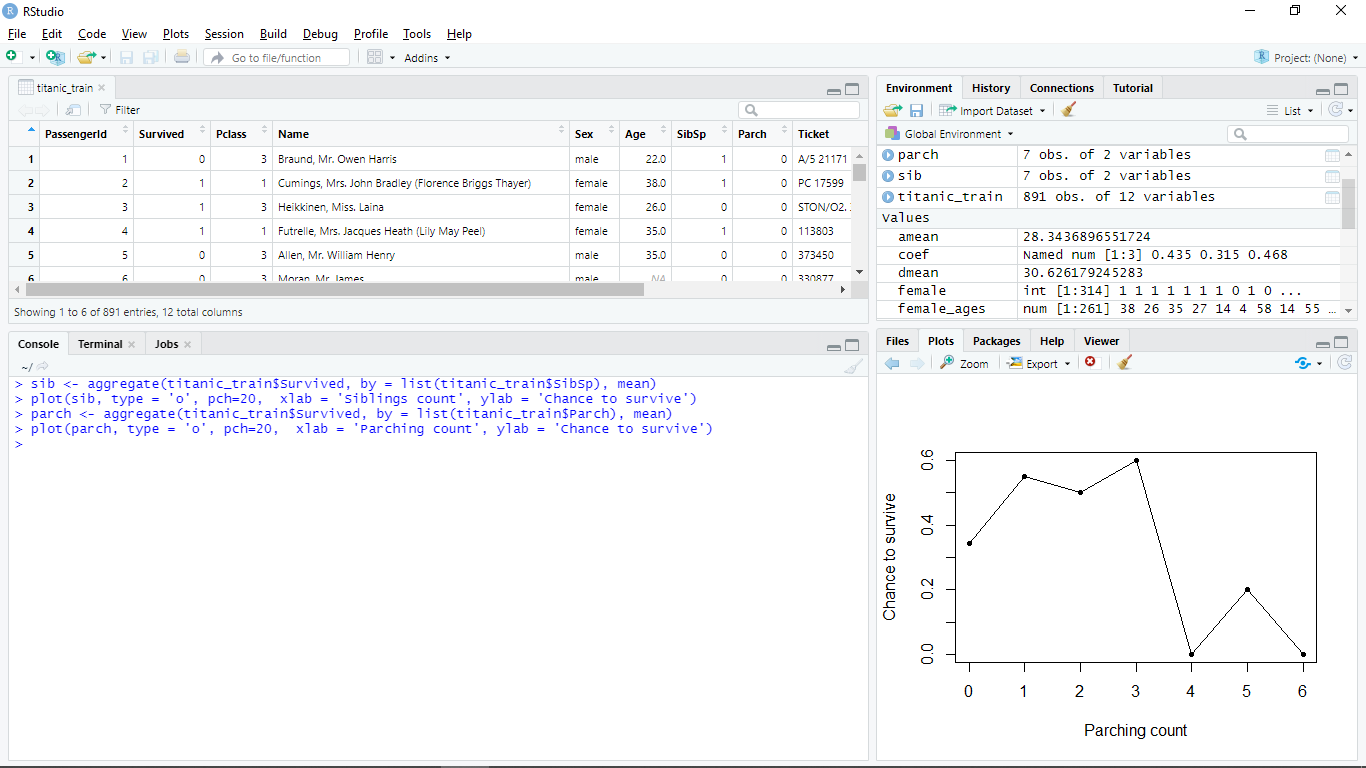
Функция *mean\_age* высчитывает средний возраст выбранного пола для каждого класса обслуживания.

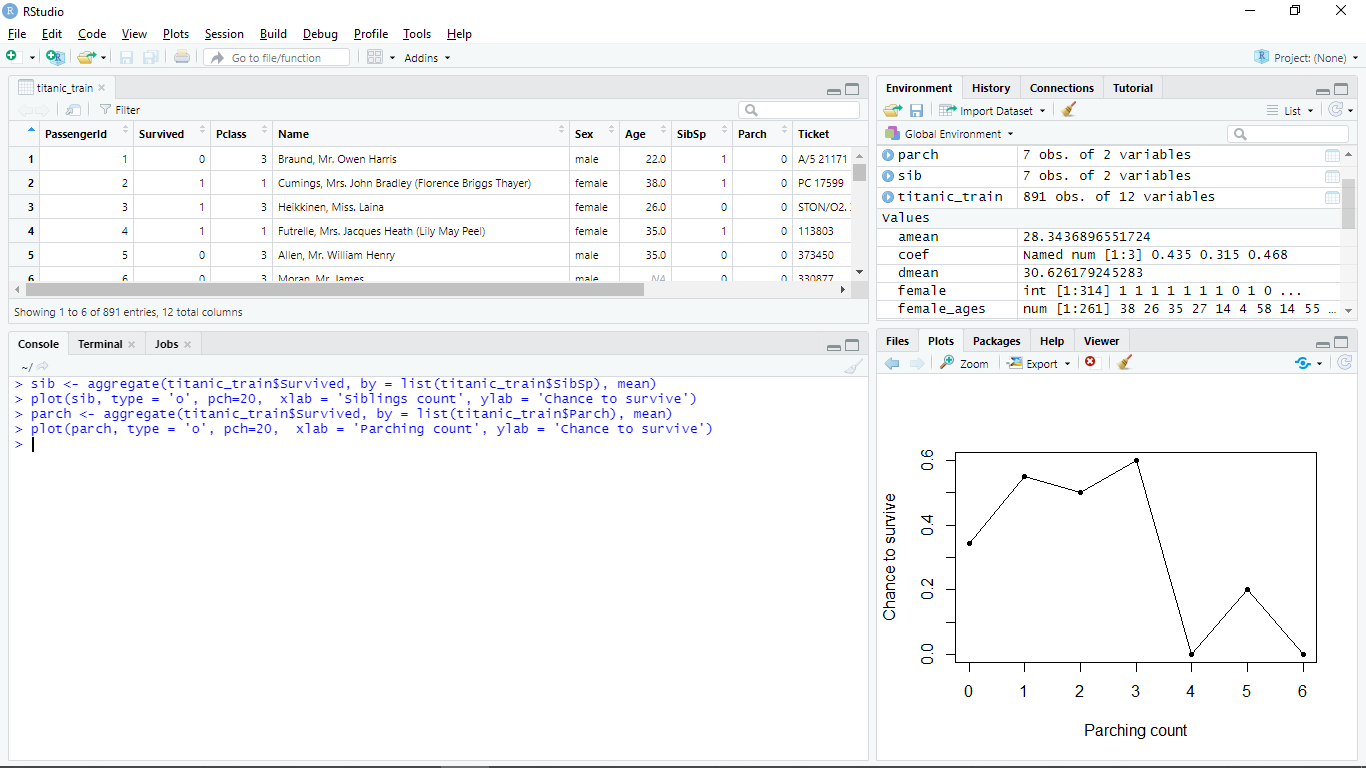
### Задание 9.

Поставленный вопрос: В каком порту на борт село больше всего женщин по отношению к общему числу пассажиров севших на борт в этом порту?

Переменная seaports хранит в себе названия городов-портов. Метод passing\_coef считает долю женщин по отношению к общему числу пассажиров, которые входили в определенном порту. После чего выводим результат. Наибольшее число женщин село в Queenstown.

### Задание 10.

 Поставленный вопрос: Правда ли что чем больше у пассажира родных на борту тем больше вероятность выжить? Проверьте эту теорию с обоими параметрами Parch и SibSp.

Первый график построен с параметром SibSp. Второй график построен с параметром Parch.

Функция aggregate разбивает базу данных на подмножества и применяем к ним функцию. Параметр by содержат набор уникальных элементов, которые способствуют для определения подмножеств.

В данном случае, мы использует столбец Survived, а в качестве параметра by выбирается столбец SibSp/Parch, в качестве функции используем функцию mean, так как требуется высчитать шанс выживания. Полученный результат выводив в качестве графика. Ось Siblings/Parching count указывает на количество родственников.