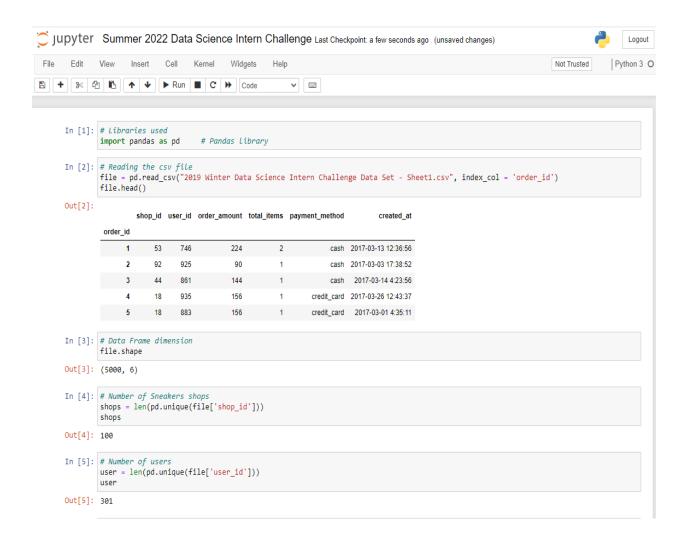
Question: 1

Given some sample data, write a program to answer the following: <u>click here to access the required data set</u>

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of \$3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

- a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
- b. What metric would you report for this dataset?
- c. What is its value?



```
In [6]: file.info()
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 5000 entries, 1 to 5000
        Data columns (total 6 columns):
                            Non-Null Count Dtype
         # Column
         0 shop_id
                             5000 non-null
             user_id
                             5000 non-null
                                             int64
             order_amount
                             5000 non-null
                                             int64
         3 total items
                             5000 non-null
                                             int64
         4 payment_method 5000 non-null 5 created_at 5000 non-null
                                             object
                                             object
        dtypes: int64(4), object(2)
        memory usage: 273.4+ KB
In [7]: # Converting the column data type
        file['created_at'] = pd.to_datetime(file['created_at'])
        file.info()
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 5000 entries, 1 to 5000
        Data columns (total 6 columns):
         # Column
                          Non-Null Count Dtype
                             5000 non-null
         0 shop_id
                                             int64
                             5000 non-null
                                             int64
         1 user_id
         2 order_amount
3 total_items
                             5000 non-null
                                             int64
                             5000 non-null
                                             int64
         4 payment_method 5000 non-null
                                             object
         5 created_at
                             5000 non-null
                                             datetime64[ns]
        dtypes: datetime64[ns](1), int64(4), object(1)
        memory usage: 273.4+ KB
```

```
In [8]: # Checking the Day window of the dataset
file.sort_values(by ='created_at')

Out[8]:
shop_id user_id order_amount total_items payment_method created_at
order_id
```

| order_id | | | | | |
|----------|----|-----|-----|---|---------------------------------|
| 1863 | 39 | 738 | 536 | 4 | cash 2017-03-01 00:08:09 |
| 1742 | 39 | 910 | 268 | 2 | cash 2017-03-01 00:10:19 |
| 3229 | 97 | 912 | 324 | 2 | cash 2017-03-01 00:14:12 |
| 1268 | 80 | 798 | 290 | 2 | credit_card 2017-03-01 00:19:31 |
| 2690 | 49 | 799 | 258 | 2 | credit_card 2017-03-01 00:22:25 |
| | | | | | |
| 2631 | 53 | 940 | 112 | 1 | credit_card 2017-03-30 23:12:13 |
| 1686 | 34 | 818 | 244 | 2 | cash 2017-03-30 23:16:10 |
| 1475 | 21 | 815 | 142 | 1 | cash 2017-03-30 23:26:54 |
| 318 | 52 | 848 | 292 | 2 | cash 2017-03-30 23:41:34 |
| 2458 | 95 | 700 | 168 | 1 | credit_card 2017-03-30 23:55:35 |

5000 rows x 6 columns

```
In [9]: # Naively Calculated average order value.
n_cal = round(sum(file['order_amount'])/file.shape[0] , 2)
print("Naively calculated average order value (AOV): $%s" % n_cal)

Naively calculated average order value (AOV): $3145.13
```

The figure of 3145.13 dollar is calculated by dividing the 'Total Order Amount' with the 'Number of Orders'. While calculation, we ignore the fact that in single order the customer could have purchased more than one item. This misjudgement led to wrong average order value.

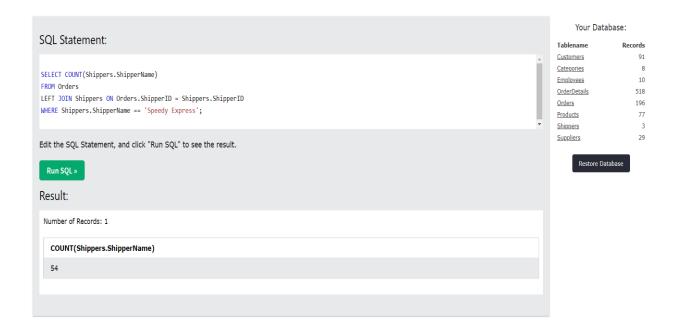
Instead of 'Number of Orders' we should have used the 'Total Number of Items' purchased by the customers and divided from 'Total Order Amount'. By making changes we get 357.92 dollars of Average Order Value which is better way to evaluate this data.

```
In [10]: # 'Total Order Amount (in 30 days)' divided by 'Total Number of Items' purchased in 30 days
a_cal = round(sum(file['order_amount'])/sum(file['total_items']),2)
print("Actual average order value (AOV): $%5"% a_cal)
Actual average order value (AOV): $357.92
```

Question: 2

For this question you'll need to use SQL. <u>Follow this link</u> to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

a. How many orders were shipped by Speedy Express in total?



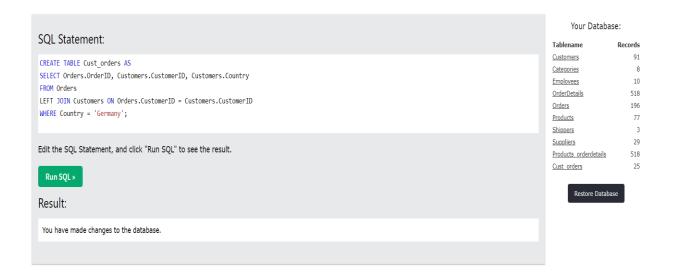
b. What is the last name of the employee with the most orders?



- c. What product was ordered the most by customers in Germany?
- 1. Table 'Products_orderdetails' created by joining tables 'OrderDetails' and 'Products'



2. Table 'Cust_orders' created by joining tables 'Orders' and 'Customers'



3. The product ordered the most by customers in Germany: Gorgonzola Telino

