# Penn State University
## SWENG 545 Data Mining
## Midterm 2

Student Name:Shashi Singh

Angel ID: sks235

**HONOR STATEMENT:**

I have completed my work according to the principle of Academic Integrity. I have neither given nor received any unauthorized aid on this assignment/examination.

Date: 04/14/2015 Initial: Shashi Singh

The following are the rules relating to this take-home exam. Any questions about interpretation of problems should be addressed to me.

1. Once you have downloaded the exam from ANGEL, you may not discuss it in any way with anyone until the exam period is over.
2. You may use a word processor or computer in preparing your answers. You may also write in pen/pencil and scan for submission.
3. You NEED NOT worry about margins, fonts, etc. on the exam as long as it is easily readable. Suggested font size is between 11 and 14 points.
4. Any violation of the rules regarding consultation with others including family members will be considered honor code violations.
5. If any problems arise during this exam, email me. I cannot make exceptions or give extensions.

## Motivation:

To learn the pattern of a typical Internet user via clustering techniques.

## Setting:

For the past seven years, the Graphics, Visualization and, Users Department of the Georgia Institute of Technology in Atlanta has conducted an international survey of World Wide Web usage as a public service in order to provide information concerning the demographics and trends of Internet access.

The goal of this assignment is to obtain a profile of the "typical" Internet user by applying clustering techniques. Such data mining task can be advantageous to e-commerce marketers so that they may tailor their advertisements to a particular people-set. These tasks may also assist to software engineers and system designers would be interested in understanding why a particular subset of the population is still uncomfortable using computers. Data set for this exam is the General Demographics dataset from the GVU WWW User Survey.*

## Exam problem

1. What are the typical groups of web users? Explain differences and similarities among groups.
2. Suggest methods of better targeting the most important customers.

### Notes:

- Your answers should be based on analyzing the provided data.
- Explain your answers and support them by results and/or screen captures.
- Try to follow the Guidelines for Data Mining Experiments from your term project.

## Scoring

Scoring will be based on:

- Appropriateness and correctness of experiment (70%)
- Discussion (20%)
- Conclusions (10%)

Particularly good discussions may result in 2% extra credit.

## Overview and Approach

The solution of this problem has been approached in the following way:

1. Data Preprocessing, which involves finding outliers, anomaly and inconsistent data.
2. Mining Structure is created with 30% train data.
3. Mining Model is created for the Microsoft Cluster to do the analysis of the data and explain the result of the analysis.

## Data Preprocessing

1. Total number of records in the Excel Sheet: 5022.
2. In order to pre-preprocessing in Excel, the Highlight Exception was run on the all the data point.

The screen-shot showing the process is given below.

**Highlight Exceptions Report for Table1**

The outlier cells are highlighted in the original table.

Exception threshold (more or fewer exceptions): 75

| Column | Outliers |
|---|---|
| Access WWW From Home | 5 |
| Access WWW From Other Places | 4 |
| Access WWW From Public Terminal | 13 |
| Access WWW From School | 18 |
| Access WWW From Work | 8 |
| Age | 12 |
| Comfort With Computers | 10 |
| Comfort With the Internet | 3 |
| Community Building | 2 |
| Country | 27 |
| Disability_Cognitive | 3 |
| Disability_Hearing | 9 |
| Disability_Motor | 3 |
| Disability_Not Impaired | 0 |
| Disability_Not Say | 4 |
| Disability_Vision | 12 |
| Education Attainment | 12 |
| Falsification of Information | 8 |
| Gender | 0 |
| Household Income | 3 |
| Kind of Area You Live In | 0 |
| Major Geographical Location | 14 |
| Marital Status | 6 |

No duplicate rows was found. Most of them are outliers, for example one record is from New Zealand because it was accessed from home daily.

3. Rename column names to add underscore in place of spaces for efficient management of pre-data processing using SQL Server. For example, “Community Membership_Family” renames to “Community_Membership_Family.”

4. Age Column: Replaced 15-Nov with 11-15 in the Age column. There are 50 records having 15-Nov in the age column.

5. Community_Membership_Family: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

6. Community_Membership_Hobbies: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

7. Community_Membership_None: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

8. Community_Membership_Other: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

9. Community_Membership_Political: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

10. Community_Membership_Professional: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

11. Community_Membership_Religious: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

12. Community_Membership_Support: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

13. Disability_Cognitive: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

14. Disability_Hearing: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

15. Disability_Motor: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

16. Disability_Not_Impaired: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

17. Disability_Not_Say: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

18. Disability_Vision: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

19. How_You_Heard_About_Survey_Banner: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

20. How_You_Heard_About_Survey_Friend: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

21. How_You_Heard_About_Survey_Mailing_List: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

22. How_You_Heard_About_Survey_Others: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

23. How_You_Heard_About_Survey_Printed_Media: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

24. How_You_Heard_About_Survey_Remebered: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

25. How_You_Heard_About_Survey_Search_Engine: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

26. How_You_Heard_About_Survey_Usenet_News: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

27. How_You_Heard_About_Survey_WWW_Page: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

28. Reasons_for_Not_Purchasing_Bad_experience: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

29. Reasons_for_Not_Purchasing_Bad_press: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

30. Reasons_for_Not_Purchasing_Cannot_find: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

31. Reasons_for_Not_Purchasing_Company_policy: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

32. Reasons_for_Not_Purchasing_Easier_locally: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

33. Reasons_for_Not_Purchasing_Enough_info: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

34. Reasons_for_Not_Purchasing_Judge_quality: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

35. Reason_for_Not_Purchasing_Never_tried: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

36. Reasons_for_Not_Purchasing_No_credit: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

37. Reasons_for_Not_Purchasing_Not_applicable: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

38. Reasons_for_Not_Purchasing_Not_home: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

39. Reasons_for_Not_Purchasing_Not_option: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

40. Reasons_for_Not_Purchasing_Other: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

41. Reasons_for_Not_Purchasing_Prefer_people: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

42. Reasons_for_Not_Purchasing_Privacy: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

43. Reasons_for_Not_Purchasing_Receipt: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

44. Reasons_for_Not_Purchasing_Security: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

45. Reasons_for_Not_Purchasing_Too_complicated: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

46. Reasons_for_Not_Purchasing_Uncomfortable: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

47. Reasons_for_Not_Purchasing_Unfamiliar_vendor: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

48. Skill_Test_Bought_Book: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

49. Skill_Test_Changed_Cookie: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

50. Skill_Test_Changed_Startup: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

51. Skill_Test_Chat: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

52. Skill_Test_Created_Page: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

53. Skill_Test_Customized_Page: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

54. Skill_Test_Major_Purchase: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

55. Skill_Test_Placed_Order: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

56. Skill_Test_Radio: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

57. Skill_Test_Telephone: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

58. Skill_Test_Took_Seminar: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

59. Skill_Test_Used_Directory: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

60. Where_Revenues_are_From_Contracts_with_government: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

61. Where_Revenues_are_From_Contracts_with_other: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

62. Where_Revenues_are_From_Contracts_with_private: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

63. Where_Revenues_are_From_Dont_know: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

64. Where_Revenues_are_From_Donations: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

65. Where_Revenues_are_From_Govt_Appropriations: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

66. Where_Revenues_are_From_Not_say: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

67. Where_Revenues_are_From_Other: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

68. Where_Revenues_are_From_Sales_to_government: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

69. Where_Revenues_are_From_Sales_to_private: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

70. Where_Revenues_are_From_User_fees: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

71. Who_Pays_for_Access_Dont_Know: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

72. Who_Pays_for_Access_Other: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

73. Who_Pays_for_Access_Parents: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

74. Who_Pays_for_Access_School: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

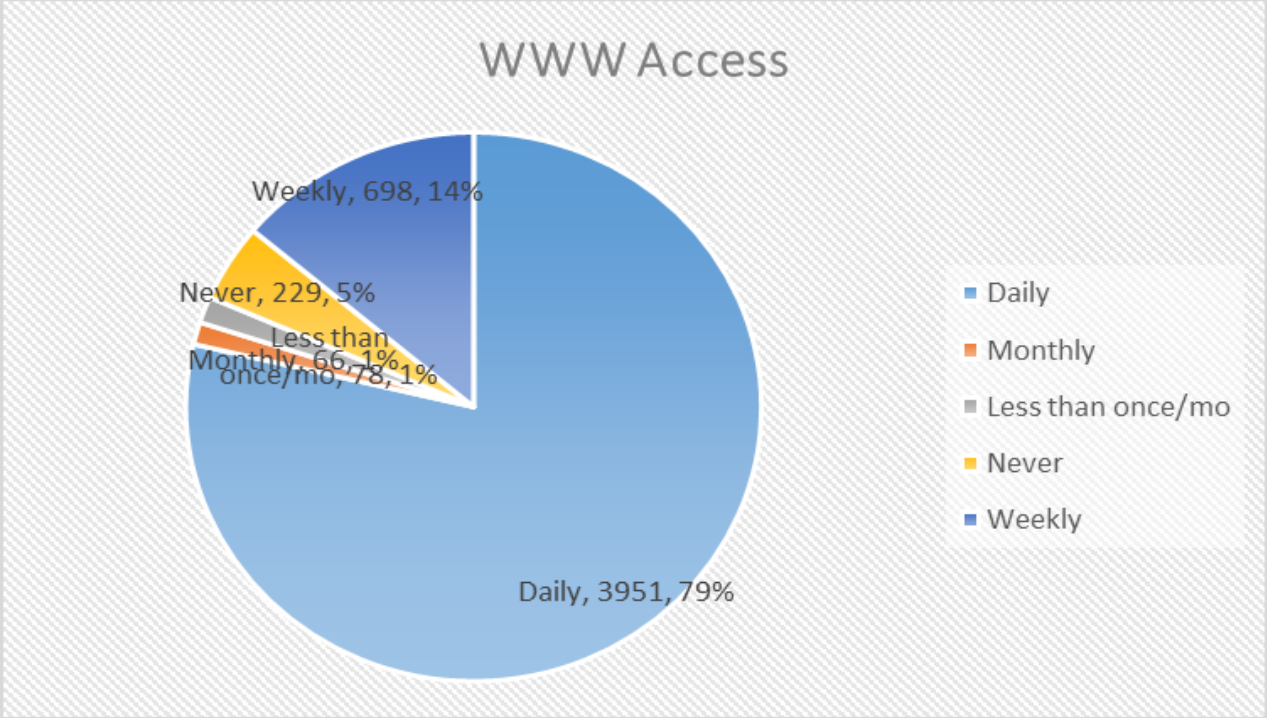
75. Who_Pays_for_Access_Self: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

76. Who_Pays_for_Access_Work: The values in this column are transformed to ‘TRUE’ for ‘1’ and ‘FALSE’ for ‘0’ to allow clustering algorithms to run.

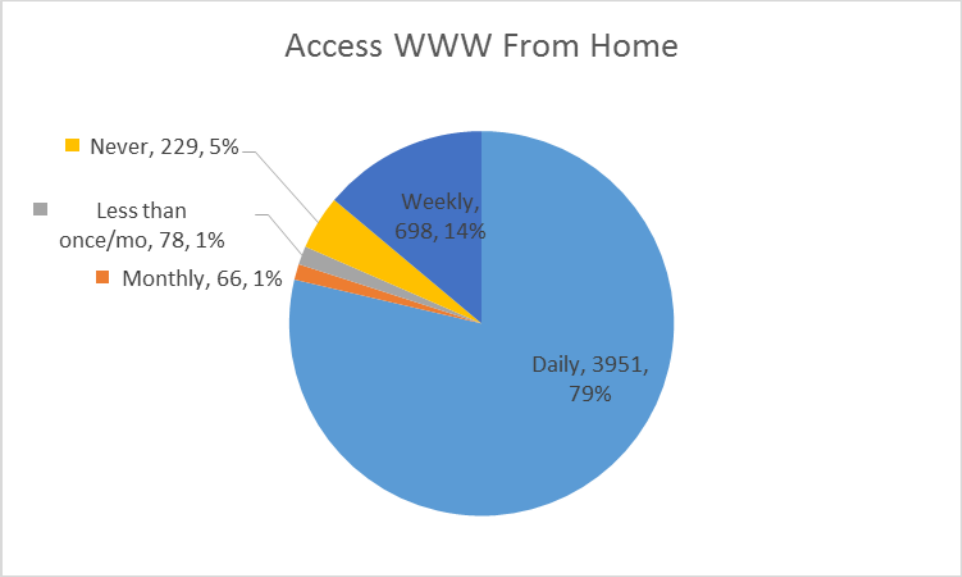
## Data Analysis

## WWW Access Frequency

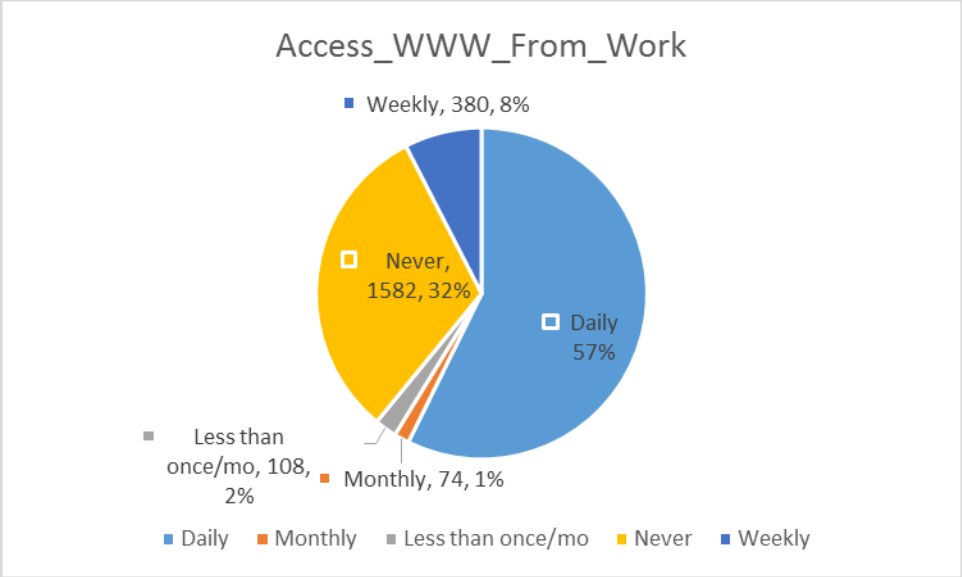
Based on the pie chart given below, 79% people access the Web daily, 14% weekly, around 1% less than once/month, around the same monthly, and around 5% population never access the web.



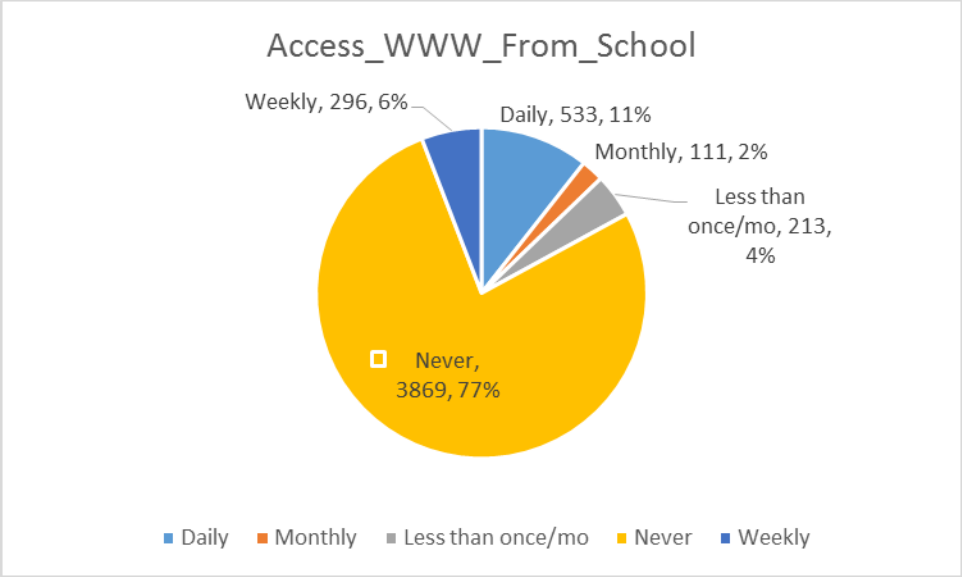
## WWW Access Frequency From Home



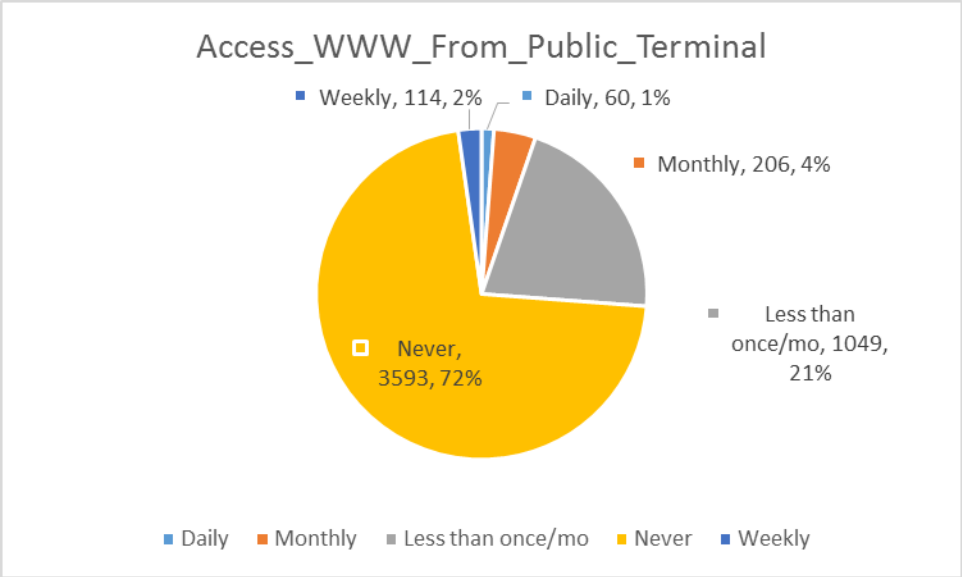

## WWW Access Frequency From Work



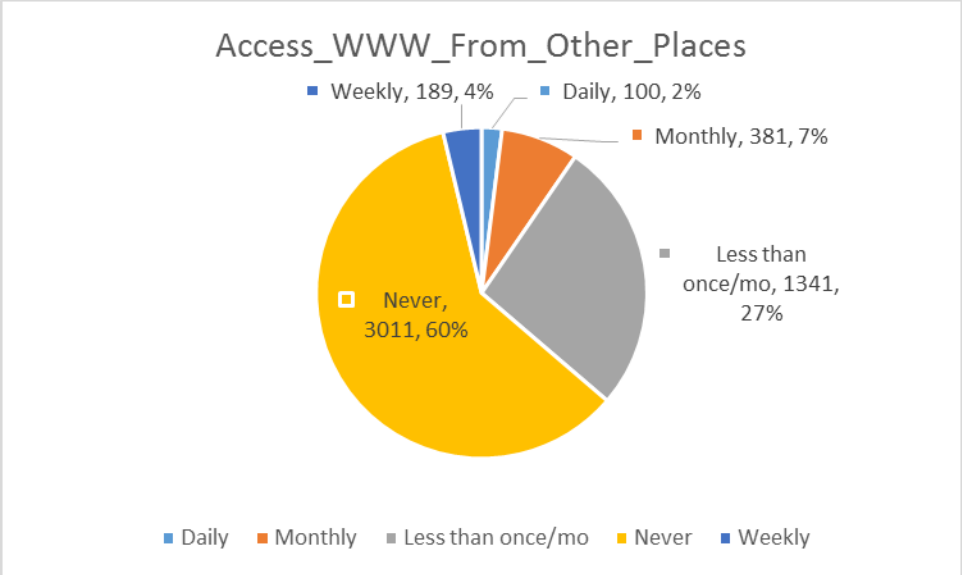
**WWW Access Frequency  From School**



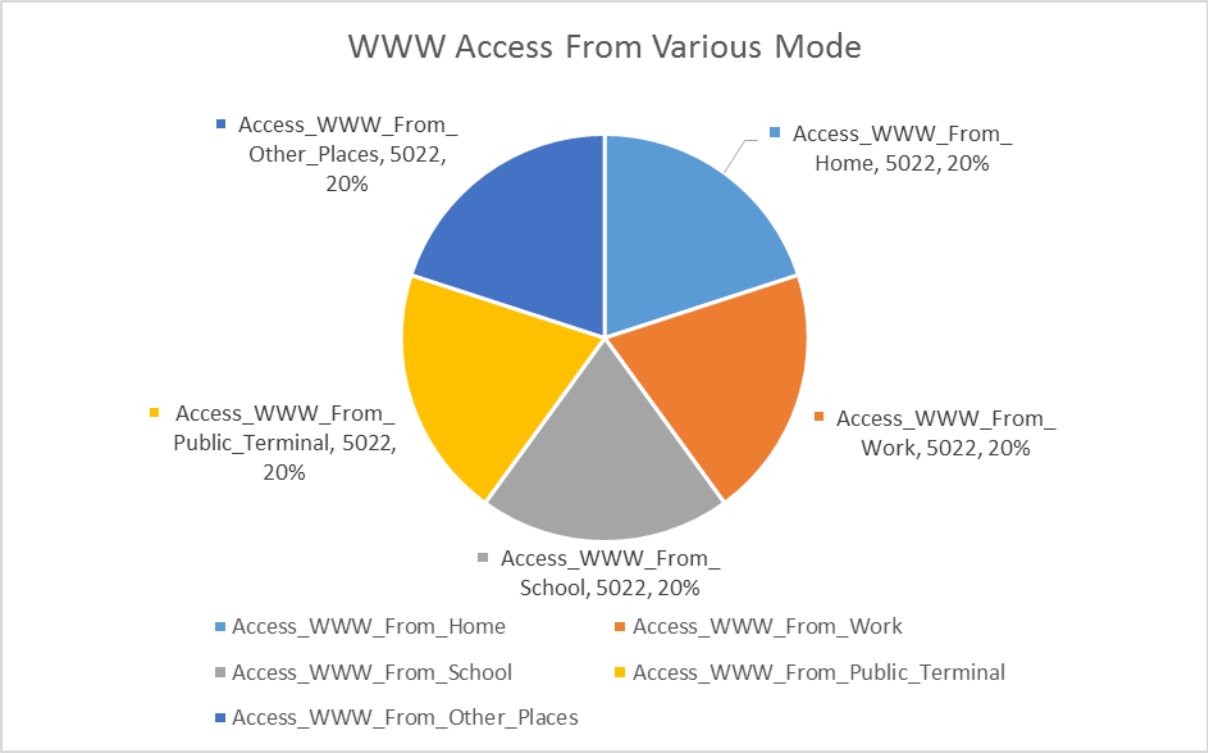
## WWW Access Frequency From Public Terminal




## WWW Access Frequency From Other Places



**WWW Access From Various Sources**



**Question 1: What are the typical groups of web users? Explain differences and similarities among groups.**

In order to answer this question, I followed the following approach based on CRISP:

1. First imported the cleaned data from SQL-Server after doing preprocessing. The preprocessing performed on the dataset is explained above.
2. **Cluster Variables**

   Following are chosen as cluster variables:

   - Access_WWW_From_Home
   - Access_WWW_From_Work
   - Access_WWW_From_School
   - Access_WWW_From_Public_Terminal
   - Age
   - Country
   - Education Attainment
   - Gender
   - Household income
   - Kind of Area you live in
   - Major geographical location
   - Married
   - Number of Children in House
   - Occupation
   - Primary Computer Platform
   - Primary Industry
   - Primary Language
   - Race
   - Sector
   - Years

   Most of the other columns have been ignored as they increase the odds of most of variable not being distinct. Some of them highly correlated as they increase the likelihood of overlapped.

3. **Choice of Algorithm**: Used K-means. This gives stable cluster with same seed. Other reason to choose K-means is have each case falling into exact segment – no overlapping as EM provides. This would give better way to target specialized group of users – without any ambiguities as there is no overlap.

4. **Number of cluster**: Choice of 4 cluster is to simple categorization so that cluster can be meaningfully differentiated as opposed many overlapping and ambiguous clusters.

5. **Different group of web users:**

   **Segment/Cluster #1:**

   Male, Access WWW from work daily, Married, Mac/Sys 8 Computing Platform,

Household income over $100K

Differences from other group:
Female, Never Access from work, Single, Student, age between 16-20 and 21-25.

**Segment/Cluster #2:**
Access from work never, sector other, retired, occupation other, years on internet under 6 month

Differences from other group
Access from work daily, years on internet 4-6 years, computing platform Mac/Sys8

**Segment/Cluster #3:**
Single, Student, age between 21-25, access from school daily, male

Differences from other group
Married, Access WWW from School Never, Female, Access from public terminal never

**Segment/Cluster #4:**
Female, access from work daily, divorced, education masters

Differences from other group
Male, Access from work Never, Student, Age 16-20

**In terms of similarities**, based on the cluster algorithm chosen – K means—there are no overlapping data sets. It means there are no similarities in terms of datasets grouped in different identified clusters. If the EM Means was chosen as a clustering algorithm, there was a probability of some datasets overlapping.

## 2. Suggest methods of better targeting the most important customers.

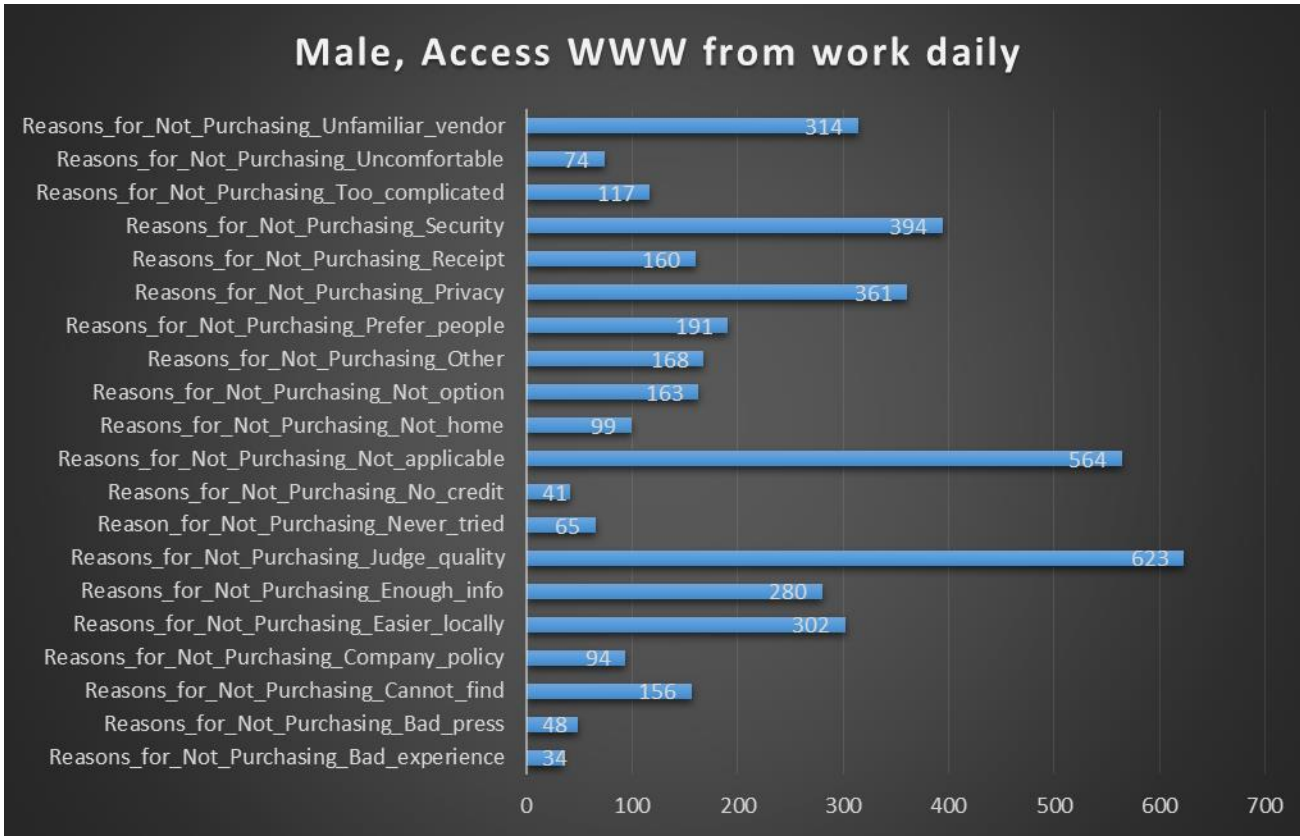
The answer to this question is given based on the dividing the data sets into 4 clusters which are identified as the followings:

- Male, Access WWW from work daily
- Single, Student
- Female, access from work daily
- Access WWW from work never, retired

### 1. Male, Access WWW from work daily

Looking at chart of this segment, some concerns can be discerned from the chart. For example, people is this group are not purchasing because of the following dominant noticeable reasons: (The reason listing is sorted: highest concern is listed first.)

- Judge quality
- Concerned about security
- Concern about privacy
- Unfamiliar vendor
- Purchasing locally is easier.

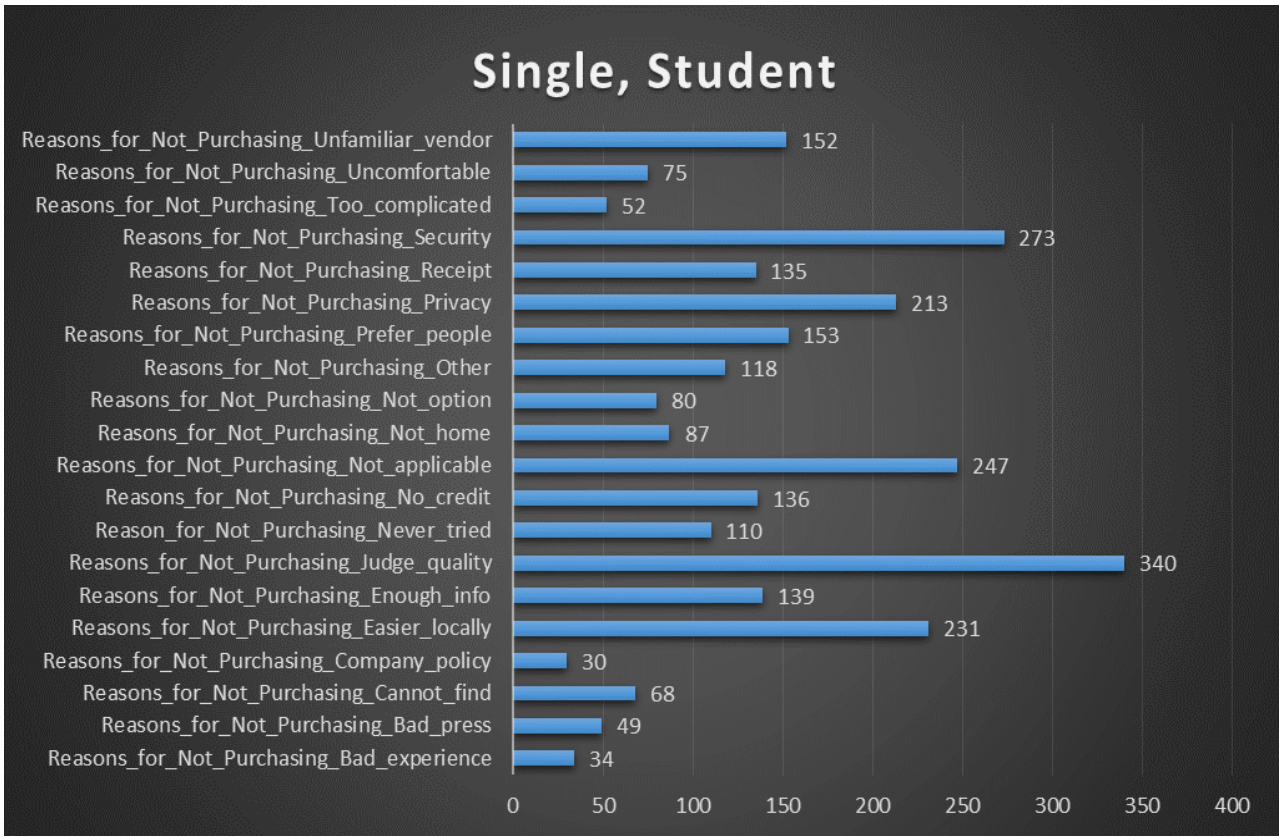


Based on the above findings, improving the quality of items. improving  the security, privacy and displaying items from the familier vendors, and making it more easier by on line – would help more people of this group buy on-line.

## 2. Single, Student

Looking at chart of this segment, some concerns can be discerned from the chart. For example, people is this group are not purchasing because of the following dominant noticeable reasons: (The reason listing is sorted: highest concern is listed first.)

- Judge quality
- Concerned about security
- Purchasing locally is easier.

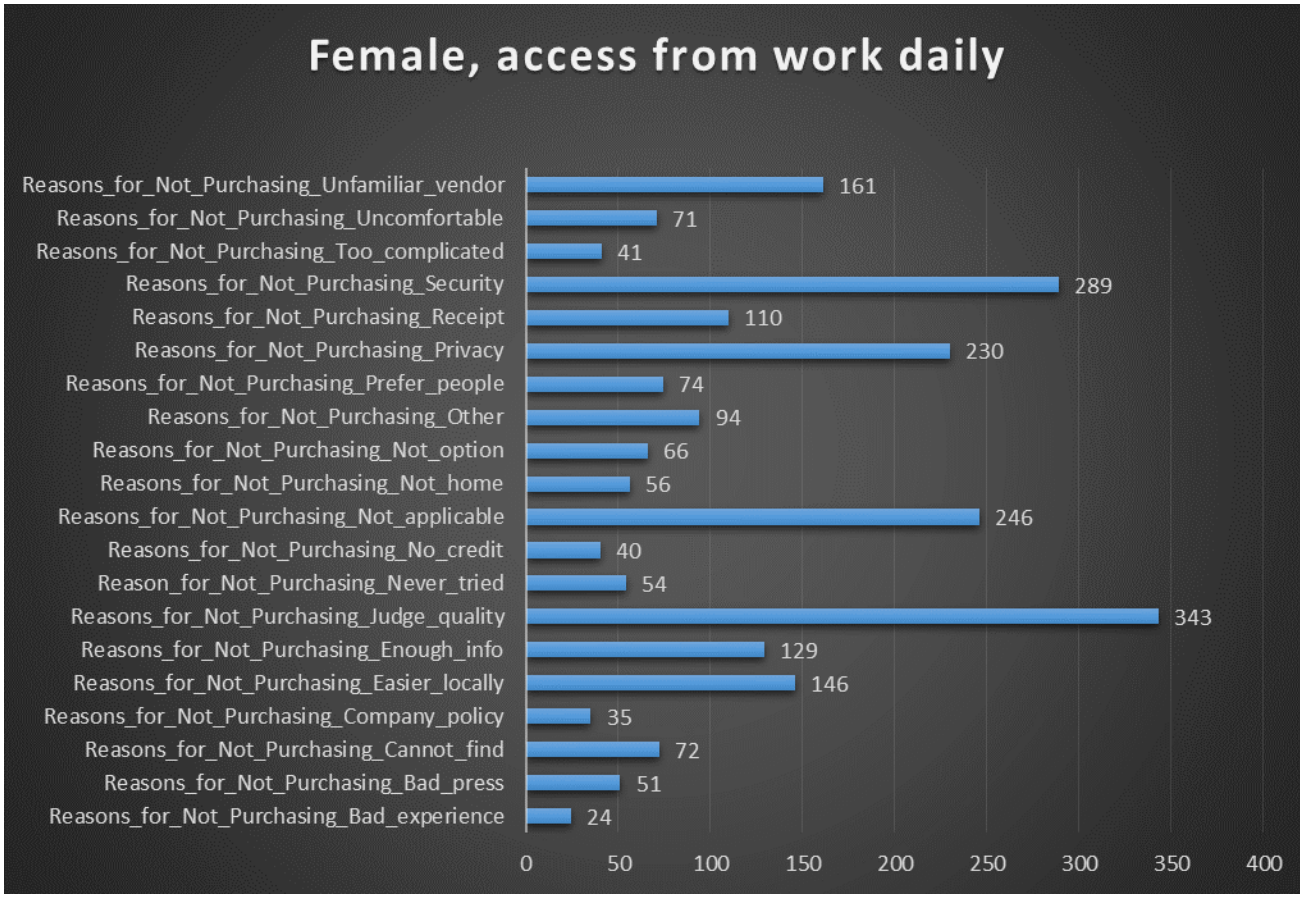



Based on the above findings, improving the quality of items. improving  the security, and making it more easier by on line – would help more people of this group buy on-line.

### 3. Female, access from work daily

Looking at chart of this segment, some concerns can be discerned from the chart. For example, people is this group are not purchasing because of the following dominant noticeable reasons: (The reason listing is sorted: highest concern is listed first.)

- Judge quality
- Concerned about security
- Concerned about privacy

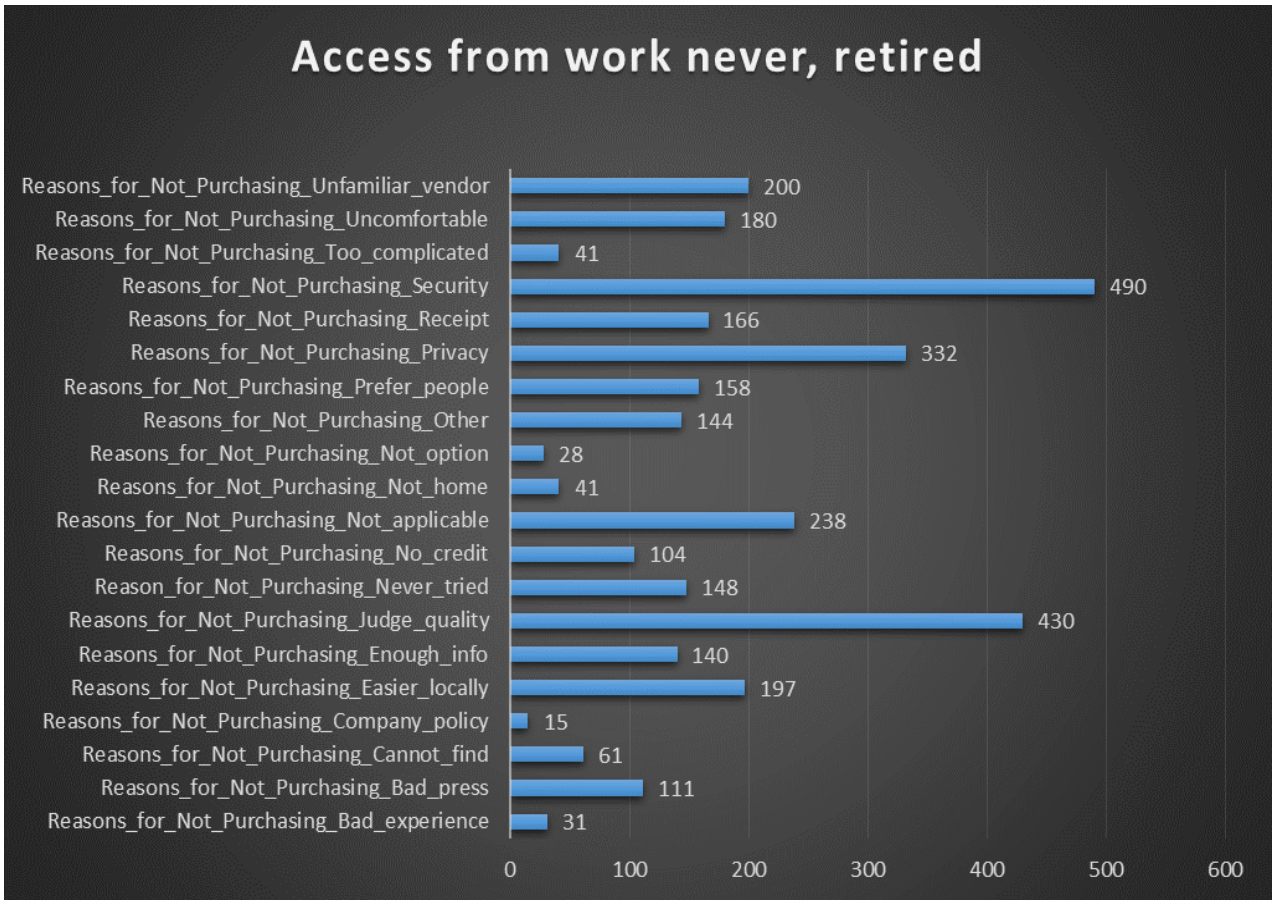



Based on the above findings, improving the quality of items. improving the security, and privacy - would help more people of this group buy on-line.

## 4. Access WWW from work never, retired

Looking at chart of this segment, some concerns can be discerned from the chart. For example, people is this group are not purchasing because of the following dominant noticeable reasons: (The reason listing is sorted: highest concern is listed first.)

- Concerned about security
- Judge quality
- Concerned about privacy




Based on the above findings, mproving the security, improving the quality of items. and privacy - would help more people of this group buy on-line.