The file **telco.xls** contains data about 1000 clients of the telecommunication company (personal client information and information about the services used by clients).

|  |  |
| --- | --- |
| **Name of the variable** | **Description of the variable** |
| region | Geographic indicator (Zone 1, Zone 2, Zone 3) |
| tenure | Months with service |
| age | Age in years |
| marital | Marital status (Unmarried, Married) |
| address | Years at current address |
| income | Household income in thousands |
| ed | Level of education (Did not complete high school, High school degree, Some college, College degree, Post-undergraduate degree) |
| employ | Years with current employer |
| retire | Retired (No, Yes) |
| gender | Gender (Male, Female) |
| reside | Number of people in household |
| tollfree | Toll free service (Yes, No) |
| equip | Equipment rental (Yes, No) |
| callcard | Calling card service (Yes, No) |
| wireless | Wireless service (Yes, No) |
| longmon | Long distance last month (duration) |
| tollmon | Toll free last month (duration) |
| equipmon | Equipment last month (duration) |
| cardmon | Calling card last month (duration) |
| wiremon | Wireless last month (duration) |
| multline | Multiple lines (Yes, No) |
| voice | Voice mail (Yes, No) |
| internet | Internet (Yes, No) |
| called | Caller ID (Yes, No) |
| callwait | Call waiting (Yes, No) |
| forward | Call forwarding (Yes, No) |
| confer | 3-way calling (Yes, No) |

The file **SMEs.xls** contains data about Small and Medium Enterprises (SMEs) in 85 federal subjects of the Russian Federation for 3 years: 2016, 2017 and 2018.

|  |  |
| --- | --- |
| **Name of the variable** | **Description of the variable** |
| id | Id of the federal subject |
| year | Calendar year |
| Region | Name of the federal subject |
| District | Name of the federal district |
| Rabotniki | Number of employees in Small and Medium Enterprises (SMEs) (people) |
| SubMSP | Number of SMEs (units) |
| Vnovsozd | Number of newly created SMEs (units) |
| oborot | Turnover of SMEs (in rubles) |
| InvestFC | Investments in fixed assets of SMEs (in rubles) |
| fedsub | Aggregated amount of federal subsidies for the development of SMEs (in rubles) |
| minekfed | Federal subsidy of the Ministry of Economic Development for SMEs (in million rubles) |
| minekreg | Regional subsidy of the Ministry of Economic Development for SMEs (in million rubles) |
| fz22344 | Financial support for SMEs for public procurement under 44 and 223 Federal Laws (in million rubles) |
| saldfinresult | Balanced financial result of SMEs (in million rubles) |
| newrabmesta | New jobs created by SMEs (thousand people) |
| newSMBdo30 | New SMEs created by entrepreneurs under 30 years old (units) |
| naselenie | Population of the federal subject of the Russian Federation (people) |
| edu30 | Number of individuals under the age of 30 who have completed training in educational programs aimed at acquiring business skills and creating SMEs (thousand people) |
| biznav | Number of SMEs registered on the SME Business Navigator Portal and supported using the SME Business Navigator services (units) |
| im | Number of SME properties in the list of state property (units) |
| infrmoney | Allocated financial resources by infrastructure organizations of the federal subject for the development of SMEs (million rubles) |
| infrrazv | Number of SME support infrastructure organizations in a federal subject of the Russian Federation (units) |
| suppmsp | Number of SMEs that received support from SME support infrastructure organizations in a federal subject of the Russian Federation (units) |
| vrp | Gross regional product of the federal subject (million rubles) |
| budg | Budget of the federal subject (million rubles) |

| **Student** | **Binary regression: Dependent variable** | **Binary regression: Predictors** | **Panel data analysis: Dependent Variable** | **Panel data analysis: Predictors** |
| --- | --- | --- | --- | --- |
| Бабаев Расул Ильхам оглы | tollfree | * region * tenure * age * marital * income   + any other 3 variables | Rabotniki | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Баталова Лиана Ильмировна | equip | * marital * age * employ * retire * gender   + any other 3 variables | SubMSP | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Белавенцев Валерий Евгеньевич | callcard | * age * marital * reside * retire * gender   + any other 3 variables | Vnovsozd | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Беляева Сабина Александровна | wireless | * income * ed * gender * reside * age   + any other 3 variables | oborot | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Бутаков Дмитрий Викторович | multline | * employ * retire * gender * age * marital   + any other 3 variables | InvestFC | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Волков Андрей Андреевич | voice | * age * marital * reside * retire * region   + any other 3 variables | fedsub | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Гадиев Михаил Искандерович | internet | * income * ed * gender * marital * age   + any other 3 variables | newSMBdo30 | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Гончаров Дмитрий Андреевич | called | * retire * gender * income * tenure * age   + any other 3 variables | minekreg | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Данилов Алексей Андреевич | callwait | * address * income * region * age * retire   + any other 3 variables | fz22344 | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Джамбонг Тенке Ханк-Дебэн - | forward | * employ * retire * age * marital * adress   + any other 3 variables | saldfinresult | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Донская Софья Алексеевна | confer | * employ * retire * age * marital * adress   + any other 3 variables | newrabmesta | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Жданов Вадим Борисович | tollfree | * income * ed * gender * marital * age   + any other 3 variables | newSMBdo30 | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Зубарев Антон Анатольевич | equip | * income * ed * gender * marital * age   + any other 3 variables | minekfed | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Калинин Антон Игоревич | callcard | * tenure * age * marital * address * employ   + any other 3 variables | edu30 | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Карелина Елена Андреевна | wireless | * region * tenure * age * marital * address   + any other 3 variables | saldfinresult | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Колчанова Алина Викторовна | multline | * region * gender * tenure * age * marital   + any other 3 variables | Biznav | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Костюченко Илья Игоревич | voice | * marital * address * income * ed * region   + any other 3 variables | Im | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Кугай Виктор Евгеньевич | internet | * marital * age * region * retire * gender   + any other 3 variables | infrmoney | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Кумар Махеш - | called | * marital * address * income * ed * region   + any other 3 variables | infrrazv | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Мохаммад Али - | callwait | * retire * gender * reside * region * age   + any other 3 variables | suppmsp | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Редникина Дарья Юрьевна | forward | * income * ed * reside * marital * age   + any other 3 variables | Rabotniki | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Савельев Семен Юрьевич | confer | * marital * address * income * age * region   + any other 3 variables | fedsub | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Соколов Семен Константинович | tollfree | * ed * employ * tenure * retire * gender   + any other 3 variables | Biznav | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Суворов Николай Михайлович | equip | * region * tenure * retire * age * gender   + any other 3 variables | oborot | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Филимонов Артём Александрович | callcard | * retire * gender * reside * region * age   + any other 3 variables | suppmsp | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |
| Шорина Анна Алексеевна | wireless | * marital * address * income * ed * region   + any other 3 variables | infrrazv | 3 continuous  variables which have the highest correlation with the dependent variable + any other 3 variables |

Please select variables for the analysis based on the information from the above table and consider the following tasks.

**Task 1 Binary regression**

* Do the statistical descriptive analysis of all the variables which will be used in binary regression analysis. Each quantitative output should be accompanied with your textual interpretation of the obtained statistics.
* Analyze all pairwise relationships between the variables which will be used in binary regression analysis based on the suitable statistical methods. Make conclusions about each obtained result.
* Do the binary regression analysis. Formulate all research hypotheses which will be verified during the data analysis.
* Specify the regression equation. Assess the goodness-of-fit of the model.
* Describe the relationships between each predictor and dependent variable in detail. Explain the meaning of each regression coefficient.
* Perform the model diagnostics.
* Make final conclusions about the results of analysis.

**Task 2 Panel Data Analysis**

* Formulate hypotheses about the existence of relationships between the dependent and independent variables.
* Conduct descriptive data analysis taking into account that the date have panel structure.
* Conduct tests to select one of three models: pulled regression model, model with fixed effects, and model with random effects. Choose the right model and justify your choice.
* Describe the resulting model as a whole.
* Describe in detail the relationships between the dependent variable and each predictor.
* Draw final conclusions from the analysis.

Please note that in task 3 and task 4 it is not supposed that you will cite or copy text from external resources. You should express your own opinion and propose your own research examples. If the text is partly or completely copied from the Internet or lectures the task will not be accepted.

**Task 3 Multinomial Regression**

Give an example of a research task that can be solved using the multinomial logistic regression method. Specify the following:

* Purpose of the study
* Research hypotheses (at least three)
* Dependent variable and its’ values
* Predictors (at least three) and their values
* What indicators need to be calculated to assess the quality of the resulting regression model? What are the values of these indicators if the model is excellent?
* What indicators need to be calculated to do the diagnostics of the regression model? What are the values of these indicators in the model is excellent?

**Task 4 Time Series Data Analysis**

* Specify the main characteristics of the time series data.
* Give an example of time series data.
* Based on this example, formulate a research problem that can be solved using time series data analysis methods.
* What problems and specific moments could complicate the time series analysis and how to cope with them?

Please send the Jupyter Notebook with comments and answers to amelikyan@hse.ru. The task should be done **individually**.The deadline for submitting the task is **28 April 16:00.**