Hal yang harus di cleaning

* Umum
  + Anggun Pratami – Error pada kolom sehingga datanya bergeser (done)
  + Yunaldi Angga Putra - Error pada kolom sehingga datanya bergeser (done)
  + Check duplicate data (done - no duplicate)
* Berapa bulan sebelum lulus Anda mencari kerja? (di luar pekerjaan sambilan)
  + Minus Value
    - There are 5 minus values that must be a mistake because they stated that they worked before graduating from the previous column. The list are:

| year\_of\_survey | nim | nama | start\_jobseek\_before |
| --- | --- | --- | --- |
| 2022 | 12115022 | Henry Harianto | -2 |
| 2022 | 15115096 | Brainy Martumbur | -2 |
| 2022 | 11915016 | Sugiarti | -1 |
| 2022 | 13515058 | Afif Bambang Prasetia | -1 |
| 2022 | 12215093 | Nevi Cahya Winofa | -3 |

* + Check Outlier
    - 243 outliers from 14716 data, it mean only 1.6% of outlier
    - Median 2.00 and mean 3.315
  + Check Data Distribution
    - The data has a positive skew
    - it’s better to use **median** rather than mean
* Berapa bulan setelah lulus Anda mencari kerja? (di luar pekerjaan sambilan)
  + Check minus value
    - The minus value doesn’t represent the mencari pekerjaan sebelum lulus. Ask your supervisor to ask a recommendation
    - Also the minus value all have the same value in column that stated they start job seeking process after they graduate
    - Almost all of them are working but 2 of them is unemployed and one of them are continue their study
  + Check Outlier
    - In 2022 there are 2 people that start a job after 4.5 and 1.5 million months?
    - The list of anomaly data are:

| year\_of\_survey | nim | nama | start\_jobseek\_after |
| --- | --- | --- | --- |
| 2022 | 15415020 | Merlin Ayu Simangunsong | 4500000 |
| 2022 | 13315035 | Andi Waluyo | 1500000 |

* + - Other than that there are no outlier (2 out of 14716 data)
  + Check Data Distribution
    - The data has negative skew/tail
    - **Median** = 1 and need to do decide the other 2 outlier to define the mean
* Manakah cara yang anda gunakan untuk mencari pekerjaan? (waiting for team internal meeting)
  + Satukan menjadi satu kolom
  + Tahun 2021 ada other (ganti kata-katanya menjadi other atau buat tabel baru dan jika ada tambahkan jadikan other
    - Done because the answer are all blanks
* Berapa perusahaan / instansi / institusi yang sudah Anda lamar sebelum memperoleh pekerjaan pertama?
  + Check minus value
    - It feels like it's impossible to have a job in minus try to some company (it’s fine if it's 0) and there are 23 data that are minus. the list are below:

| year\_of\_survey | nim | nama | jobseek\_n |
| --- | --- | --- | --- |
| 2021 | 10214071 | I Wayan Nadiantara | -2 |
| 2021 | 15514039 | Ilham Muhamad Nurmansyah | -2 |
| 2021 | 15414042 | Putu Indah Adnyani | -1 |
| 2021 | 13714019 | Ahmad Miftahul Anwar | -10 |
| 2021 | 15414038 | Zahara Sitta Iskandar | -3 |
| 2021 | 18214044 | Atika Rahmawati Yuliantoputri | -1 |
| 2021 | 12214064 | Azahra Refani | -1 |
| 2021 | 12114096 | Hartman Mesaroziduhu Zebua Wanggai | -1 |
| 2021 | 18214022 | Mokhamad Arfan Wicaksono | -1 |
| 2021 | 13114033 | Wandi Prasetia | -2 |
| 2021 | 12014026 | Aidina Anisa | -2 |
| 2022 | 13515014 | Mokhammad Ferdi Ghozali | -1 |
| 2022 | 18315002 | Galih Fajar Ramadhan | -1 |
| 2022 | 15115096 | Brainy Martumbur | -50 |
| 2022 | 13015035 | Indah Dwi Rachmawati | -1 |
| 2022 | 14515027 | Vindy Rahmadhanita Ridwan | -1 |
| 2022 | 15315004 | Kinanti Aldhia Naura | -1 |
| 2022 | 11215039 | Astrid Theola Azarine | -4 |
| 2022 | 10615055 | Audie Masola Putra | -2 |
| 2022 | 15415028 | Muhammad Rizki Rayani Ramadani | -5 |
| 2022 | 15415029 | Elistya Hestin Kusumarini | -1 |
| 2022 | 15215021 | Mahira Raihana Putri | -1 |
| 2022 | 11515029 | Zayinatun Biladiyah Al khoeriyah | -1 |

* + Check the higher such as 100 or above because it’s diluar nurul
  + Check Outlier
    - It has 3 outlier from current data before handling minus or diluar nurul data (2 from 9999 and 1 from 99999)
    - The list are:

| year\_of\_survey | nim | nama | jobseek\_n |
| --- | --- | --- | --- |
| 2021 | 13314028 | Nita Tri Nugraheni | 9999 |
| 2022 | 11415013 | Irfan Imaduddin | 99999 |
| 2022 | 11415010 | Mureen Dengganni Samosir | 9999 |

* + Check Data DIstribution
    - The data are negatively skew with the tail are on the left side
    - Median = 4.00 and Mean = 20.29
    - Mean are shifting to the right because of the anomaly data of 9999 and 99999
* Gambarkan pekerjaan yang ideal
  + Satukan menjadi satu kolom (waiting for the next team internal meeting)
  + Other jadikan other dan buat satu kolom baru untuk alasannya
    - Done because there is only one data that fills this column.
* Berapa bulan waktu yang digunakan (sebelum kelulusan) untuk memperoleh pekerjaan pertama?
  + Check kembali apakah 48+ bulan sebelum graduate itu mungkin? liat angkatan dan tahun survey atau kelulusan
  + Check Outlier
    - there are 67 outlier out of 14716 data (0.45% of outliers)
    - Median = 2.00 and Mean = 3.315
  + Check Data Distribution
    - The data is positively skew and its better to use **median** rather than mean
* Berapa bulan waktu yang digunakan (sesudah kelulusan) untuk memperoleh pekerjaan pertama?
  + Check again apakah 48+ bulan untuk mendapatkan pekerjaan setelah graduation itu possible? liat kapan lulusnya dan liat apakah dia melanjutkan ke jenjang pendidikan berikutnya atau tidak?
  + Check Outlier
    - There are 366 outlier out of 14716 data (2.48% data are outlier)
    - Median = 1 and Mean = 835
  + Check Data Distribution
    - The data is positively skewed it means use **median** instead of mean
* Melalui apa Anda mendapatkan pekerjaan pertama?
  + Check is there any anomaly in the data or outside the category (done)
* Jika relasi, siapakah relasinya?
  + Check is there any anomaly in the data or outside the category (done)
* Berdasarkan persepsi anda, penialian penting bagi Perusahaan?
  + Satukan menjadi satu kolom (waiting for the next internal meeting)
  + Other jadikan Other dan penjelasannya terdapat di kolom baru (done)
* Berapa perusahaan / instansi / institusi yang telah Anda masuki untuk bekerja (termasuk perusahaan sendiri) sejak lulus dari perguruan tinggi?
  + Check Minus Value
    - It’s impossible to have a minus value for this question because the minimum frequency of company is 0
    - Therefore we need action for this data either change it to median or pretend it was a typing mistake therefore we will convert the minus value to positive value
  + Check outlier
    - From all of the data, there is only 1 outlier. That 1 outlier can be considered as Anommaly:

| year\_of\_survey | nim | nama | experience\_n |
| --- | --- | --- | --- |
| 2019 | 13212073 | Muhammad Gemilang Rilvi | 771111 |

* + - Another form of anomaly data is changing jobs so fast. Example i tried to calculate the from moment they start college to the year of conducted survey and see how often they changing their job with the formula is if the fastest changing career is once every 3 month and it leaves 4 times switching career every year. so (year\_of\_survey - angkatan) \* 4 < total switching company (experience\_n) and if that true we can called anomaly and here are the list:

| year\_of\_survey | angkatan | nim | nama | experience\_n |
| --- | --- | --- | --- | --- |
| 2018 | 2011 | NA | Feisal Nawwaf | 100 |
| 2018 | 2011 | NA | Fanny Dwikartini | 200 |
| 2018 | 2011 | NA | Grahadi Pandu Mughny | 35 |
| 2018 | 2011 | NA | Bima Ulma | 30 |
| 2018 | 2011 | NA | Irfan Nasrullah | 123 |
| 2018 | 2011 | NA | Akbar Januari Fadlih | 30 |
| 2018 | 2011 | NA | Muhammad Arfan Pratama | 70 |
| 2018 | 2011 | NA | Jodi Pradipta | 100 |
| 2019 | 2012 | 10512010 | Tifaniar Yunitasari | 50 |
| 2019 | 2012 | 13212073 | Muhammad Gemilang Rilvi | 771111 |
| 2019 | 2012 | 13212045 | Dela Natalia Malau | 60 |
| 2019 | 2012 | 15112035 | Dzulkhi Yudha Pratama | 100 |
| 2019 | 2012 | 12212033 | Indah Cita Cahyani | 50 |
| 2020 | 2013 | 10513058 | Lusi Oksri Dona | 100 |
| 2020 | 2013 | 12813018 | Intan Hildayana | 30 |
| 2020 | 2013 | 11413041 | Katiana Apriyani Nurjanah | 35 |
| 2020 | 2013 | 10713025 | Moch Fajar Nurfadhillah Iqbal | 30 |
| 2020 | 2013 | 12013026 | Rakhadiaz Adrian Fauzan | 30 |
| 2021 | 2014 | 12114003 | Muhamad Alief Akbar | 100 |
| 2021 | 2014 | 13714007 | Egi Setiawan | 100 |
| 2021 | 2014 | 10514036 | Luttya Monica Andriani | 30 |
| 2021 | 2014 | 15314034 | Farah Nabila Vianda | 30 |
| 2021 | 2014 | 13614015 | Ahmad Fadlilah Muzzamil | 30 |
| 2021 | 2014 | 15014066 | Rizky Andini Putri | 50 |
| 2021 | 2014 | 15314058 | Sania Septiani Mulyawan | 200 |
| 2021 | 2014 | 10514006 | Cucu Fitria Lestari | 80 |
| 2021 | 2014 | 13714042 | Fairuz Surya Annabil | 100 |
| 2021 | 2014 | 12914007 | Aziz Nurul Hidayat | 100 |
| 2021 | 2014 | 12014063 | Salman Alfarisi | 100 |
| 2021 | 2014 | 19014153 | Reza Maulana Firdaus | 40 |
| 2021 | 2014 | 11514047 | Gina Afdilla | 31 |
| 2022 | 2015 | 10115068 | Kurnia Khairunnisaa | 30 |
| 2022 | 2015 | 10115048 | Rizki Teguh Kurniawan | 30 |
| 2022 | 2015 | 11215035 | Harryyanto Ishaq Agasi | 100 |
| 2022 | 2015 | 13715044 | Yuni Eva Kartika | 50 |
| 2022 | 2015 | 10415034 | Fajar Dewantoro | 50 |
| 2022 | 2015 | 15415004 | Tengku Nathasya | 200 |
| 2022 | 2015 | 13115092 | Muhammad Khazmy | 50 |
| 2022 | 2015 | 15115056 | Roni Junianto Sitompul | 50 |
| 2022 | 2015 | 11915003 | Hasan Nashrullah | 100 |
| 2022 | 2015 | 12015035 | Mohammad Firman Wahyudi | 30 |
| 2022 | 2015 | 10415031 | Andrew Daniel Koesoemo Sianturi | 60 |
| 2022 | 2015 | 11415009 | Anggun Nurul Fatimah | 50 |

* + Check data distribution
    - The data are having positively skew and the mean are affected by the outlier so it’s better to use **median** rather than mean
    - Median = 2.0 and Mean = 68.1
* Dalam seminggu, berapa jam rata-rata Anda bekerja?
  + Check minus value
    - Workhour per week cannot be minus therefore there must be a mistake when they input the number and we recommend changing its value to positive. Here are the 3 negative value:

| year\_of\_survey | nim | nama | workhour |
| --- | --- | --- | --- |
| 2021 | 13414004 | Ni Luh Putu Lilis Sinta Setiawati | -13 |
| 2021 | 17314022 | Hanifah Rachmat Fairuz Lubis | -3 |
| 2022 | 14515016 | Gabriela Mega Rahayu | -1 |

* + - But be careful when changing some of the data because 3 hours a week in a private company is questionable. But that 1 hour a week in non profit can be reasonable.
  + Small value under 10 hours, see if it real or something like enterpreuneur
    - First, I would like to see under 10 hours of work a week and alumni that work in a company (“bekerja” status) and not in an internship program. There are 666 alumni that fall into this category. In our opinion this can be the mistake of the alumni not reading the work hours as weekly hours. So we recommend to either make it times 5-6 (Monday to Friday or Saturday) or don’t insert it to a calculation. We dont recommend it to make all the data to be daily since it is only 4.5% of total data.
  + Check high value of working hour
    - We define that the maximum work hour in a day is 10 hours (normal without overtime) and work 6 days with sunday is a holiday so the total work hour a week is 60 hours a week and it's only for alumni who work in a company. For those who didn't work in the company either as an entrepreneur or entrepreneur we get the maximum hour of 24 hours per day so 168 hours a week and the value above 168 hours are considered as anomaly data. There are 269 anomaly data in that data according to that hypothesis.
  + Check Data Distribution
    - The data has slightly positively skewed with median score 40.00 and mean score 72.09.
* Sejak
  + Ubah menjadi date format (half done)
    - There is a problem because there are 118 columns that have only a year in that given date. Please ask your supervisor for what to do with that data because if we just add month to that data it will affecting the analysis
  + Clean the data so that all of the data have the same date format (done)
  + Putuskan apakah menggunakan dd-mm-yyyy atau mm-yyyy (liat buku pdf untuk guidenya dan tentukan monthnya apakah menggunakan angka atau kata-kata)
    - menggunakan format dd-mm-yyyy (done)
* Komentar negatif tentang perusahan (opsional if there is time or needed data)
  + Makna dari nilai 0 apakah tidak ada atau kesalahan penulisan karna ada 2 orang
* Bidang usaha bekerja
  + Ada data yang diluar kategori coba di check dan di cleaning
  + Terdapat 2 data yang harus di bersihkan coba check apa yang harus dilakukan
* Nama Kantor
  + Apa makna dari 0? Ubahkah menjadi null?
  + Apa makna dari 1? Apakah diubah menjadi null?
  + Ubah – dan – menjadi null value saja jika status alumni tersebut wirausaha jika bekerja maka ubah menjadi null
* Gaji perbulan
  + Check anomaly gaji
    - according to this website <https://www.cnbcindonesia.com/lifestyle/20230510090432-33-436050/10-pekerjaan-dengan-gaji-tertinggi-di-indonesia-2023> alumni who has salary more than Rp. 80.000.000,- and work in indonesia as a worker (“bekerja” status) is categorize as anomaly, but according to the result there are no anomaly in the data.
    - And from the monthly salary that alumni have below Rp. 500.000,- there are no anomalies in the data.
  + Checking the data distribution
    - Median = Rp. 7.200.000 and Mean = Rp. 9.008.067
    - Therefore we can visualize both or just use median because the data is slightly positively skewed
* Bonus pertahun
  + Terdapat data yang bergeser kolomnya tolong di check lagi (done)
  + Check the data distribution
    - The data has slightly positively skewed so you can either put both or show only the median.
    - Median = Rp. 16.000.000,- and Mean = Rp. 28.630.000,-
* Apakah pekerjaan yang Anda lakukan di tempat bekerja sesuai dengan bidang kuliah?
  + Check ya dan tidaknya dan pastikan match dengan status pekerjaan dia apakah bekerja, wirausaha, tidak bekerja, dsb (done)
    - After the code there are 1 name that appear to said yes to align with their education background but not working, but after thorough investigation its all clear
* Apakah pekerjaan yang Anda lakukan di tempat wirausaha sesuai dengan bidang kuliah?
  + Check ya dan tidaknya dan pastikan match dengan status pekerjaan dia apakah bekerja, wirausaha, tidak bekerja, dsb
    - There are 7 data that don't match because they are working but also said that they wirausaha column said yes. 7 of them said yes to both questions about working and entrepreneurship. Therefore we decide to delete the entrepreneurship answer and only get their answer from their working question. Here are 7 data:

| year\_of\_survey | nim | nama | kesesuaian\_wirausaha |
| --- | --- | --- | --- |
| 2018 |  | Syaf Musyaffaq | ya |
| 2018 |  | Syah Adi Setiadi | ya |
| 2018 |  | Taufiq Akbari Utomo | ya |
| 2018 |  | Tessa Sarah Olystia | ya |
| 2019 | 11212016 | Hamdin Kifahul Muhajir | ya |
| 2019 | 10712034 | Shella Rahmayanti | ya |
| 2019 | 15012156 | Aulia Alfan Nugraha | ya |

* Apa alasan Anda memutuskan berwirausaha? (Jawaban bisa lebih dari 1) [Cita-cita]
  + Gabungkan menjadi satu kolom (waiting for the next team internal meeting)
  + Bagian describe jadikan other dan tambahkan satu kolom untuk menjelaskan describe tersebut (done)
* Jenis usaha yang Anda kembangkan saat ini? (Jawaban bisa lebih dari 1) [Usaha sendiri / tidak memiliki pegawai]
  + Gabungkan menjadi satu kolom (waiting for the next team internal meeting)
  + Bagian describe jadikan other dan tambahkan satu kolom untuk menjelaskan describe tersebut (done)
* Bulan dan tahun Anda melanjutkan studi
  + Data type ubah ke date (half done)
  + Date format ubah menjadi mm-yyyy (Samakan dengan format date sebelumya apakah menggunakan angka atau sebutkan bulannya June, august, dll) (done mm-yyyy)
  + Check anomaly
    - Responden are writing it wrong

| year\_of\_survey | nim | nama | grad\_start |
| --- | --- | --- | --- |
| 2022 | 10215027 | Daniel Suprobo Wangsaputra | Juli / Agustus 2021 |
| 2022 | 15115001 | Muhammad Fakhriy Ramadhan | 2 tahun |

* + - Sistem error (25 data):

| year\_of\_survey | nim | nama | grad\_start |
| --- | --- | --- | --- |
| 2021 | 15514019 | Muhammad Luqman Al Hakim | Jan-70 |
| 2021 | 13114132 | Genesia Wahyu Saputro | Jan-70 |
| 2021 | 18214005 | Aditya Ghifari | Jan-70 |
| 2021 | 10614020 | Muhammad Rifki Ramadhan | Jan-70 |
| 2021 | 13114025 | Bivynka Kemala Damanik | Jan-70 |
| 2021 | 15014073 | Irvin Nathaniel Tobing | Jan-70 |
| 2021 | 10114010 | Mudita Gunawan | Jan-70 |
| 2021 | 13114033 | Wandi Prasetia | Jan-70 |
| 2021 | 17314011 | Hazimah Ulfah Az Zaky | Jan-70 |
| 2021 | 15014058 | Indira Ayu Revita Putri | Jan-70 |
| 2021 | 15014014 | Vallian Yuda Wiratama | Jan-70 |
| 2021 | 13514065 | Geraldi Dzakwan | Jan-70 |
| 2021 | 15814026 | Gabriel Ami Yansyah | Jan-70 |
| 2021 | 15314030 | Windy Chyntia Dewi | Jan-70 |
| 2021 | 13214077 | Hansen | Jan-70 |
| 2021 | 15714026 | Dameria Maranatha Gloriani | Jan-70 |
| 2021 | 13514033 | Faza Thirafi | Jan-70 |
| 2021 | 17314010 | Tsamara Alifia | Jan-70 |
| 2021 | 10514079 | Ratih Tyas Widara | Jan-70 |
| 2021 | 10314002 | Muhammad Rezky | Jan-70 |
| 2021 | 11214020 | Tjokorda Isteri Indira Sekar Wangi Pudak | Jan-70 |
| 2021 | 15014089 | Marvin Gunawan | Jan-70 |
| 2021 | 15214076 | Raynaldo Tantawi Swardi | Jan-70 |
| 2021 | 13214115 | Fitri Nadira Adiputri | Jan-70 |
| 2021 | 13314088 | Putu Handre Kertha Utama | Jan-70 |

* + - Respondent only type the year without month:

| year\_of\_survey | nim | nama | grad\_start |
| --- | --- | --- | --- |
| 2022 | 13615047 | Diksan Muhammad | 2020 |
| 2022 | 13215046 | Benita Angela Titalim | 2020 |
| 2022 | 11215036 | Diah Ayu Prawitasari | 2022 |

* Berapa lama Anda bekerja setelah lulus (pekerjaan pertama) sebelum memutuskan untuk berhenti atau tidak bekerja lagi? (dalam bulan)
  + All of the data is blank or empty (done)