1. Have a data source and cleaning notes appendix: cover how you checked the quality of data.
   1. Checked number of employees in PAYE vs number of jobs. But this is an apples to oranges comparison. However,
      1. PAYE ~29m workers
      2. ASHE ~20m "jobs": 15m is full time jobs, 6m is part time jobs.
      3. If we consider on average all part-time workers do 0.431 of full time hours then number of workers in ASHE = 15m + 6m/0.5 = ~27m. This comes very close to the PAYE number of ~29m workers. The difference is perhaps because I assumed the wrong average of 0.5.

1 I got the 0.43 value from here: <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/timeseries/ybvb>. If full time hours ~37, and if on average a part-time worker, works 16 hours, then 16/37 = 0.43

* 1. “ASHE is based on a 1% sample of jobs taken from HM Revenue and Customs' Pay As You Earn (PAYE) records. Consequently, individuals with more than one job may appear in the sample more than once.” (from the ASHE excel sheet).

PAYE is collected every month. ASHE is collected annually. This means that wages from both of them are not directly comparable. PAYE monthly wages are calculated as hourly wage\*365/12 (refer to PAYE methodology document). ASHE annual wage is calculated as “Annual estimates are provided for the tax year that ended on 5th April in the reference year.” Moreover, “Hourly and weekly estimates are provided for the pay period that included a specified date in April. They relate to employees on adult rates of pay, whose earnings for the survey pay period were not affected by absence.”

Therefore, to make both of them comparable, I should take gross hourly wage in April. Perhaps, I can go beyond April in the main analysis, but I should note that April is the common month.

1. Have a methodology appendix.
2. Have some robustness checks appendix (or in the main body): For instance, in the wage data by occupation, we have some estimates that are not reliable. Talk about what happens to analysis when we drop them vs when we don’t.
3. Since this is an explainer for the public, explain why we have the median and mean.

Model:

1. There are two industries, Hi and Li, high paying and low paying industry respectively.
2. There are two occupations within each industry, Ho and Lo, high and low paying occupation respectively.
3. We observe only whether an occupation is high or low paying. Not which industry this is in. Perhaps for our analysis there might be a clever mathematical transformation that would return some useful results even if we don’t observe this. For instance, if running a regression of wage on age, occupation, we could simply control for industry at the age level.
4. Within each occupation, there are two age groups, young and old.
5. In the short run, we’re concerned about demographic transition between high and low paying industries and/or occupations.
6. Why might we be concerned about this? Consider a few examples where inflation due to a wage-price spiral, comes from whether someone invests or saves:
   1. Old people in high paying jobs in high paying industry, drops out of work force after covid, and takes early retirement. We wonder whether this job is filled by a younger person? This could imply two mechanisms for a wage-price induced inflation. Both of them are determined by the degree to which individuals choose to save or invest:
      1. On the one hand, if it is a high-tech/R&D type industry, then either the large number of young would mean some structural shift in innovation, and R&D that could enter the market in the future. The pace of R&D determines the pace of prices, and thus endogenously determines wage again, and so on.
      2. On the other hand, if it is a low R&D type industry, then there are a lot of young with high wages, who are not investing into new tech, i.e., they are not saving, and rather consumption increases, thus increasing prices, and so on.
   2. Young people in low paying jobs in low paying industry, could move to low paying job in high paying industry. Again inflation could arise out of a number of mechanisms. The same savings-investment mechanisms as above could kick in, but the size of it wouldn’t matter.

These are just two examples. There could be a number of possibilities arising in the data. To be precise, we have 2 (increase or decrease in wage) \* 2 (increase or decrease in hours worked – a measure of productivity) \* 2 (whether someone is young or old) \* 2 (high or low occupation) \* 2 (high or low industry) \* 2 (high or low share of individuals in each category, i.e., given increase in wage, old, high occupation, high industry, are the share of individuals in this category large enough to matter?) = 64 possibilities. Here we considered a simplistic model that has only two possibilities, the analysis would perhaps get tricky if we have a continuum rather than categories (for ex: 100 industries ranked by wage etc.). Or perhaps it won’t get complicated – maybe some kind of basic regressions might fix this issue. Another issue is that each possibility could result in several mechanisms (save or invest; good management or bad management), and therefore the stories that we could tell increase exponentially. So a key question is should we worry about mechanisms? And if so, which all? I think we should come back to these last 2 questions, after we do point 7 and find results from there.

1. Going back in time, do we observe the same possibilities in the data from a time series perspective? Point 6 was easy to do because we just needed to look at the transition from 2021 to 2022. So for each of the 32 possibilities, we would have only 2 data points, one in 2021 and another in 2022. I think if there weren’t any structural breaks then we could simply run a regression and see the transition from t to t+1 on average. Possible contenders for structural breaks are the 2007 recession (solution: just use data from 2009 onwards), or Paris agreement and other climate kind of policies that might have majorly shifted factors of production. The latter might not necessarily break our “model”. For instance, suppose fossil fuel based automobile industry was high paying in time t. In the next time period, automobile just went down in ranking, and was replaced by electric automobile industry. Given this ranking, if share of individuals by occupation and age group remains constant, then we are not worried. We can make a very easy comparison between patterns before and after covid. Given ranking, if share changes, then we need to see (and think about) what these changes are, and how it relates to the changes we observe between 2021 and 2022.